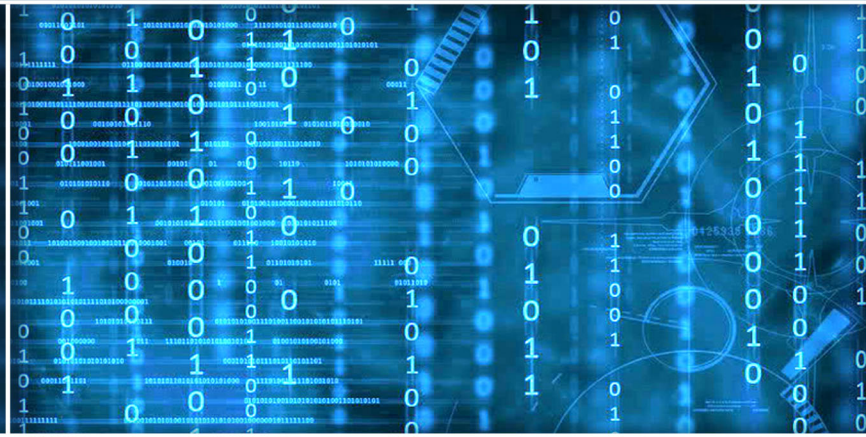


Volume 7 Issue 3

March 2016



ISSN 2156-5570(Online)

ISSN 2158-107X(Print)



[www.ijacsa.thesai.org](http://www.ijacsa.thesai.org)

# Editorial Preface

## *From the Desk of Managing Editor...*

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

**Thank you for Sharing Wisdom!**

**Managing Editor**  
**IJACSA**  
**Volume 7 Issue 3 March 2016**  
**ISSN 2156-5570 (Online)**  
**ISSN 2158-107X (Print)**  
**©2013 The Science and Information (SAI) Organization**



# Editorial Board

## Editor-in-Chief

**Dr. Kohei Arai - Saga University**

*Domains of Research: Technology Trends, Computer Vision, Decision Making, Information Retrieval, Networking, Simulation*

---

## Associate Editors

**Chao-Tung Yang**

**Department of Computer Science, Tunghai University, Taiwan**

*Domain of Research: Software Engineering and Quality, High Performance Computing, Parallel and Distributed Computing, Parallel Computing*

**Elena SCUTELNICU**

**"Dunarea de Jos" University of Galati, Romania**

*Domain of Research: e-Learning, e-Learning Tools, Simulation*

**Krassen Stefanov**

**Professor at Sofia University St. Kliment Ohridski, Bulgaria**

*Domains of Research: e-Learning, Agents and Multi-agent Systems, Artificial Intelligence, Big Data, Cloud Computing, Data Retrieval and Data Mining, Distributed Systems, e-Learning Organisational Issues, e-Learning Tools, Educational Systems Design, Human Computer Interaction, Internet Security, Knowledge Engineering and Mining, Knowledge Representation, Ontology Engineering, Social Computing, Web-based Learning Communities, Wireless/ Mobile Applications*

**Maria-Angeles Grado-Caffaro**

**Scientific Consultant, Italy**

*Domain of Research: Electronics, Sensing and Sensor Networks*

**Mohd Helmy Abd Wahab**

**Universiti Tun Hussein Onn Malaysia**

*Domain of Research: Intelligent Systems, Data Mining, Databases*

**T. V. Prasad**

**Lingaya's University, India**

*Domain of Research: Intelligent Systems, Bioinformatics, Image Processing, Knowledge Representation, Natural Language Processing, Robotics*

## Reviewer Board Members

- **Aamir Shaikh**
- **Abbas Al-Ghaili**  
Mendeley
- **Abbas Karimi**  
Islamic Azad University Arak Branch
- **Abdelghni Lakehal**  
Université Abdelmalek Essaadi Faculté  
Polydisciplinaire de Larache Route de Rabat, Km 2 -  
Larache BP. 745 - Larache 92004. Maroc.
- **Abdul Razak**
- **Abdul Karim ABED**
- **Abdur Rashid Khan**  
Gomal University
- **Abeer Elkorany**  
Faculty of computers and information, Cairo
- **ADEMOLA ADESINA**  
University of the Western Cape
- **Aderemi A. Atayero**  
Covenant University
- **Adi Maaita**  
ISRA UNIVERSITY
- **Adnan Ahmad**
- **Adrian Branga**  
Department of Mathematics and Informatics,  
Lucian Blaga University of Sibiu
- **agana Becejski-Vujaklija**  
University of Belgrade, Faculty of organizational
- **Ahmad Saifan**  
yarmouk university
- **Ahmed Boutejdar**
- **Ahmed AL-Jumaily**  
Ahlia University
- **Ahmed Nabih Zaki Rashed**  
Menoufia University
- **Ajantha Herath**  
Stockton University Galloway
- **Akbar Hossain**
- **Akram Belghith**  
University Of California, San Diego
- **Albert S**  
Kongu Engineering College
- **Alcinia Zita Sampaio**  
Technical University of Lisbon
- **Alexane Bouënard**  
Sensopia
- **ALI ALWAN**  
International Islamic University Malaysia
- **Ali Ismail Awad**  
Luleå University of Technology
- **Alicia Valdez**
- **Amin Shaqrah**  
Taibah University
- **Amirrudin Kamsin**
- **Amitava Biswas**  
Cisco Systems
- **Anand Nayyar**  
KCL Institute of Management and Technology,  
Jalandhar
- **Andi Wahyu Rahardjo Emanuel**  
Maranatha Christian University
- **Anews Samraj**  
Mahendra Engineering College
- **Anirban Sarkar**  
National Institute of Technology, Durgapur
- **Anthony Isizoh**  
Nnamdi Azikiwe University, Awka, Nigeria
- **Antonio Formisano**  
University of Naples Federico II
- **Anuj Gupta**  
IKG Punjab Technical University
- **Anuranjan misra**  
Bhagwant Institute of Technology, Ghaziabad, India
- **Appasami Govindasamy**
- **Arash Habibi Lashkari**  
University Technology Malaysia(UTM)
- **Aree Mohammed**  
Directorate of IT/ University of Sulaimani
- **ARINDAM SARKAR**  
University of Kalyani, DST INSPIRE Fellow
- **Aris Skander**  
Constantine 1 University
- **Ashok Matani**  
Government College of Engg, Amravati
- **Ashraf Owis**  
Cairo University
- **Asoke Nath**

St. Xaviers College(Autonomous), 30 Park Street,  
Kolkata-700 016

- **Athanasios Koutras**
- **Ayad Ismaeel**  
Department of Information Systems Engineering-  
Technical Engineering College-Erbil Polytechnic  
University, Erbil-Kurdistan Region- IRAQ
- **Ayman Shehata**  
Department of Mathematics, Faculty of Science,  
Assiut University, Assiut 71516, Egypt.
- **Ayman EL-SAYED**  
Computer Science and Eng. Dept., Faculty of  
Electronic Engineering, Menofia University
- **Babatunde Opeoluwa Akinkunmi**  
University of Ibadan
- **Bae Bossoufi**  
University of Liege
- **BALAMURUGAN RAJAMANICKAM**  
Anna university
- **Balasubramanie Palanisamy**
- **BASANT VERMA**  
RAJEEV GANDHI MEMORIAL COLLEGE, HYDERABAD
- **Basil Hamed**  
Islamic University of Gaza
- **Basil Hamed**  
Islamic University of Gaza
- **Bhanu Prasad Pinnamaneni**  
Rajalakshmi Engineering College; Matrix Vision  
GmbH
- **Bharti Waman Gawali**  
Department of Computer Science & information T
- **Bilian Song**  
LinkedIn
- **Binod Kumar**  
JSPM's Jayawant Technical Campus, Pune, India
- **Bogdan Belean**
- **Bohumil Brtnik**  
University of Pardubice, Department of Electrical  
Engineering
- **Bouchaib CHERRADI**  
CRMEF
- **Brahim Raouyane**  
FSAC
- **Branko Karan**
- **Bright Keswani**  
Department of Computer Applications, Suresh Gyan  
Vihar University, Jaipur (Rajasthan) INDIA
- **Brij Gupta**

University of New Brunswick

- **C Venkateswarlu Sonagiri**  
JNTU
- **Chanashekhhar Meshram**  
Chhattisgarh Swami Vivekananda Technical  
University
- **Chao Wang**
- **Chao-Tung Yang**  
Department of Computer Science, Tunghai  
University
- **Charlie Obimbo**  
University of Guelph
- **Chee Hon Lew**
- **Chien-Peng Ho**  
Information and Communications Research  
Laboratories, Industrial Technology Research  
Institute of Taiwan
- **Chun-Kit (Ben) Ngan**  
The Pennsylvania State University
- **Ciprian Dobre**  
University Politehnica of Bucharest
- **Constantin POPESCU**  
Department of Mathematics and Computer  
Science, University of Oradea
- **Constantin Filote**  
Stefan cel Mare University of Suceava
- **CORNELIA AURORA Gyorödi**  
University of Oradea
- **Cosmina Ivan**
- **Cristina Turcu**
- **Dana PETCU**  
West University of Timisoara
- **Daniel Albuquerque**
- **Dariusz Jakóbczak**  
Technical University of Koszalin
- **Deepak Garg**  
Thapar University
- **Devena Prasad**
- **DHAYA R**
- **Dheyaa Kadhim**  
University of Baghdad
- **Djilali IDOUGH**  
University A.. Mira of Bejaia
- **Dong-Han Ham**  
Chonnam National University
- **Dr. Arvind Sharma**



- Aryan College of Technology, Rajasthan Technology University, Kota
- **Duck Hee Lee**  
Medical Engineering R&D Center/Asan Institute for Life Sciences/Asan Medical Center
  - **Elena SCUTELNICU**  
"Dunarea de Jos" University of Galati
  - **Elena Camossi**  
Joint Research Centre
  - **Eui Lee**  
Sangmyung University
  - **Evgeny Nikulchev**  
Moscow Technological Institute
  - **Ezekiel OKIKE**  
UNIVERSITY OF BOTSWANA, GABORONE
  - **Fahim Akhter**  
King Saud University
  - **FANGYONG HOU**  
School of IT, Deakin University
  - **Faris Al-Salem**  
GCET
  - **Firkhan Ali Hamid Ali**  
UTHM
  - **Fokrul Alom Mazarbhuiya**  
King Khalid University
  - **Frank Ibikunle**  
Botswana Int'l University of Science & Technology (BIUST), Botswana
  - **Fu-Chien Kao**  
Da-Y eh University
  - **Gamil Abdel Azim**  
Suez Canal University
  - **Ganesh Sahoo**  
RMRIMS
  - **Gaurav Kumar**  
Manav Bharti University, Solan Himachal Pradesh
  - **George Pecherle**  
University of Oradea
  - **George Mastorakis**  
Technological Educational Institute of Crete
  - **Georgios Galatas**  
The University of Texas at Arlington
  - **Gerard Dumancas**  
Oklahoma Baptist University
  - **Ghalem Belalem**  
University of Oran 1, Ahmed Ben Bella
  - **gherabi noreddine**
  - **Giacomo Veneri**  
University of Siena
  - **Giri Babu**  
Indian Space Research Organisation
  - **Govindarajulu Salendra**
  - **Grebenisan Gavril**  
University of Oradea
  - **Gufan Ahmad Ansari**  
Qassim University
  - **Gunaseelan Devaraj**  
Jazan University, Kingdom of Saudi Arabia
  - **GYÖRÖDI ROBERT STEFAN**  
University of Oradea
  - **Hadj Tadjine**  
IAV GmbH
  - **Haewon Byeon**  
Nambu University
  - **Haiguang Chen**  
ShangHai Normal University
  - **Hamid Alinejad-Rokny**  
The University of New South Wales
  - **Hamid AL-Asadi**  
Department of Computer Science, Faculty of Education for Pure Science, Basra University
  - **Hamid Mukhtar**  
National University of Sciences and Technology
  - **Hany Hassan**  
EPF
  - **Harco Leslie Henic SPITS WARNARS**  
Bina Nusantara University
  - **Hariharan Shanmugasundaram**  
Associate Professor, SRM
  - **Harish Garg**  
Thapar University Patiala
  - **Hazem I. El Shekh Ahmed**  
Pure mathematics
  - **Hemalatha SenthilMahesh**
  - **Hesham Ibrahim**  
Faculty of Marine Resources, Al-Mergheb University
  - **Himanshu Aggarwal**  
Department of Computer Engineering
  - **Hongda Mao**  
Hossam Faris
  - **Huda K. AL-Jobori**  
Ahlia University
  - **Imed JABRI**

- **iss EL OUADGHIRI**
- **Iwan Setyawan**  
Satya Wacana Christian University
- **Jacek M. Czerniak**  
Casimir the Great University in Bydgoszcz
- **Jai Singh W**
- **JAMAIAH HAJI YAHAYA**  
NORTHERN UNIVERSITY OF MALAYSIA (UUM)
- **James Coleman**  
Edge Hill University
- **Jatinderkumar Saini**  
Narmada College of Computer Application, Bharuch
- **Javed Sheikh**  
University of Lahore, Pakistan
- **Jayaram A**  
Siddaganga Institute of Technology
- **Ji Zhu**  
University of Illinois at Urbana Champaign
- **Jia Uddin Jia**  
Assistant Professor
- **Jim Wang**  
The State University of New York at Buffalo,  
Buffalo, NY
- **John Sahlin**  
George Washington University
- **JOHN MANOHAR**  
VTU, Belgaum
- **JOSE PASTRANA**  
University of Malaga
- **Jui-Pin Yang**  
Shih Chien University
- **Jyoti Chaudhary**  
high performance computing research lab
- **K V.L.N.Acharyulu**  
Bapatla Engineering college
- **Ka-Chun Wong**
- **Kamatchi R**
- **Kamran Kowsari**  
The George Washington University
- **KANNADHASAN SURIYAN**
- **Kashif Nisar**  
Universiti Utara Malaysia
- **Kato Mivule**
- **Kayhan Zrar Ghafoor**  
University Technology Malaysia
- **Kennedy Okafor**  
Federal University of Technology, Owerri
- **Khalid Mahmood**  
IEEE
- **Khalid Sattar Abdul**  
Assistant Professor
- **Khin Wee Lai**  
Biomedical Engineering Department, University  
Malaya
- **Khurram Khurshid**  
Institute of Space Technology
- **KIRAN SREE POKKULURI**  
Professor, Sri Vishnu Engineering College for  
Women
- **KITIMAPORN CHOOCHOTE**  
Prince of Songkla University, Phuket Campus
- **Krasimir Yordzhev**  
South-West University, Faculty of Mathematics and  
Natural Sciences, Blagoevgrad, Bulgaria
- **Krassen Stefanov**  
Professor at Sofia University St. Kliment Ohridski
- **Labib Gergis**  
Misr Academy for Engineering and Technology
- **LATHA RAJAGOPAL**
- **Lazar Stošić**  
College for professional studies educators  
Aleksinac, Serbia
- **Leanos Maglaras**  
De Montfort University
- **Leon Abdillah**  
Bina Darma University
- **Lijian Sun**  
Chinese Academy of Surveying and
- **Ljubomir Jerinic**  
University of Novi Sad, Faculty of Sciences,  
Department of Mathematics and Computer Science
- **Lokesh Sharma**  
Indian Council of Medical Research
- **Long Chen**  
Qualcomm Incorporated
- **M. Reza Mashinchi**  
Research Fellow
- **M. Tariq Banday**  
University of Kashmir
- **madjid khalilian**
- **majzoob omer**
- **Mallikarjuna Doodipala**  
Department of Engineering Mathematics, GITAM  
University, Hyderabad Campus, Telangana, INDIA

- **Manas deep**  
Masters in Cyber Law & Information Security
- **Manju Kaushik**
- **Manoharan P.S.**  
Associate Professor
- **Manoj Wadhwa**  
Echelon Institute of Technology Faridabad
- **Manpreet Manna**  
Director, All India Council for Technical Education,  
Ministry of HRD, Govt. of India
- **Manuj Darbari**  
BBD University
- **Marcellin Julius Nkenlifack**  
University of Dschang
- **Maria-Angeles Grado-Caffaro**  
Scientific Consultant
- **Marwan Alseid**  
Applied Science Private University
- **Mazin Al-Hakeem**  
LFU (Lebanese French University) - Erbil, IRAQ
- **Md Islam**  
sikkim manipal university
- **Md. Bhuiyan**  
King Faisal University
- **Md. Zia Ur Rahman**  
Narasaraopeta Engg. College, Narasaraopeta
- **Mehdi Bahrami**  
University of California, Merced
- **Messaouda AZZOUZI**  
Ziane Achour University of Djelfa
- **Milena Bogdanovic**  
University of Nis, Teacher Training Faculty in Vranje
- **Miriampally Venkata Raghavendra**  
Adama Science & Technology University, Ethiopia
- **Mirjana Popovic**  
School of Electrical Engineering, Belgrade University
- **Miroslav Baca**  
University of Zagreb, Faculty of organization and  
informatics / Center for biometrics
- **Moeiz Miraoui**  
University of Gafsa
- **Mohamed Eldosoky**
- **Mohamed Ali Mahjoub**  
Preparatory Institute of Engineer of Monastir
- **Mohamed Kaloup**
- **Mohamed El-Sayed**  
Faculty of Science, Fayoum University, Egypt
- **Mohamed Najeh LAKHOUA**  
ESTI, University of Carthage
- **Mohammad Ali Badamchizadeh**  
University of Tabriz
- **Mohammad Jannati**
- **Mohammad Alomari**  
Applied Science University
- **Mohammad Haghighat**  
University of Miami
- **Mohammad Azzeh**  
Applied Science university
- **Mohammed Akour**  
Yarmouk University
- **Mohammed Sadgal**  
Cadi Ayyad University
- **Mohammed Al-shabi**  
Associate Professor
- **Mohammed Hussein**
- **Mohammed Kaiser**  
Institute of Information Technology
- **Mohammed Ali Hussain**  
Sri Sai Madhavi Institute of Science & Technology
- **Mohd Helmy Abd Wahab**  
University Tun Hussein Onn Malaysia
- **Mokhtar Beldjehem**  
University of Ottawa
- **Mona Elshinawy**  
Howard University
- **Mostafa Ezziyani**  
FSTT
- **Mouhammd sharari alkasassbeh**
- **Mourad Amad**  
Laboratory LAMOS, Bejaia University
- **Mueen Uddin**  
University Malaysia Pahang
- **MUNTASIR AL-ASFOOR**  
University of Al-Qadisiyah
- **Murphy Choy**
- **Murthy Dasika**  
Geethanjali College of Engineering & Technology
- **Mustapha OUJAOURA**  
Faculty of Science and Technology Béni-Mellal
- **MUTHUKUMAR SUBRAMANYAM**  
DGCT, ANNA UNIVERSITY
- **N.Ch. Iyengar**  
VIT University
- **Nagy Darwish**



Department of Computer and Information Sciences,  
Institute of Statistical Studies and Researches, Cairo  
University

- **Najib Kofahi**  
Yarmouk University
- **Nan Wang**  
LinkedIn
- **Natarajan Subramanyam**  
PES Institute of Technology
- **Natheer Gharaibeh**  
College of Computer Science & Engineering at  
Yanbu - Taibah University
- **Nazeeh Ghatasheh**  
The University of Jordan
- **Nazeeruddin Mohammad**  
Prince Mohammad Bin Fahd University
- **NEERAJ SHUKLA**  
ITM UNiversity, Gurgaon, (Haryana) India
- **Neeraj Tiwari**
- **Nestor Velasco-Bermeo**  
UPFIM, Mexican Society of Artificial Intelligence
- **Nidhi Arora**  
M.C.A. Institute, Ganpat University
- **Nilanjan Dey**
- **Ning Cai**  
Northwest University for Nationalities
- **Nithyanandam Subramanian**  
Professor & Dean
- **Noura Aknin**  
University Abdelamlek Essaadi
- **Obaida Al-Hazaimeh**  
Al- Balqa' Applied University (BAU)
- **Oliviu Matei**  
Technical University of Cluj-Napoca
- **Om Sangwan**
- **Omaima Al-Allaf**  
Asesstant Professor
- **Osama Omer**  
Aswan University
- **Ouchtati Salim**
- **Ousmane THIARE**  
Associate Professor University Gaston Berger of  
Saint-Louis SENEGAL
- **Paresh V Virparia**  
Sardar Patel University
- **Peng Xia**  
Microsoft

- **Ping Zhang**  
IBM
- **Poonam Garg**  
Institute of Management Technology, Ghaziabad
- **Prabhat K Mahanti**  
UNIVERSITY OF NEW BRUNSWICK
- **PROF DURGA SHARMA ( PHD)**  
AMUIT, MOEFDRE & External Consultant (IT) &  
Technology Tansfer Research under ILO & UNDP,  
Academic Ambassador for Cloud Offering IBM-USA
- **Purwanto Purwanto**  
Faculty of Computer Science, Dian Nuswantoro  
University
- **Qifeng Qiao**  
University of Virginia
- **Rachid Saadane**  
EE departement EHTP
- **Radwan Tahboub**  
Palestine Polytechnic University
- **raed Kanaan**  
Amman Arab University
- **Raghuraj Singh**  
Harcourt Butler Technological Institute
- **Rahul Malik**
- **raja boddu**  
LENORA COLLEGE OF ENGINEERNG
- **Raja Ramachandran**
- **Rajesh Kumar**  
National University of Singapore
- **Rakesh Dr.**  
Madan Mohan Malviya University of Technology
- **Rakesh Balabantaray**  
IIIT Bhubaneswar
- **Ramani Kannan**  
Universiti Teknologi PETRONAS, Bandar Seri  
Iskandar, 31750, Tronoh, Perak, Malaysia
- **Rashad Al-Jawfi**  
Ibb university
- **Rashid Sheikh**  
Shri Aurobindo Institute of Technology, Indore
- **Ravi Prakash**  
University of Mumbai
- **RAVINA CHANGALA**
- **Ravisankar Hari**  
CENTRAL TOBACCO RESEARCH INSTITUE
- **Rawya Rizk**  
Port Said University

- **Reshmy Krishnan**  
Muscat College affiliated to Stirling University.U
- **Ricardo Vardasca**  
Faculty of Engineering of University of Porto
- **Ritaban Dutta**  
ISSL, CSIRO, Tasmania, Australia
- **Rowayda Sadek**
- **Ruchika Malhotra**  
Delhi Technological University
- **Rutvij Jhaveri**  
Gujarat
- **SAADI Slami**  
University of Djelfa
- **Sachin Kumar Agrawal**  
University of Limerick
- **Sagarmay Deb**  
Central Queensland University, Australia
- **Said Ghoniemy**  
Taif University
- **Sandeep Reddivari**  
University of North Florida
- **Sanskriti Patel**  
Charotar University of Science & Technology,  
Changa, Gujarat, India
- **Santosh Kumar**  
Graphic Era University, Dehradun (UK)
- **Sasan Adibi**  
Research In Motion (RIM)
- **Satyena Singh**  
Professor
- **Sebastian Marius Rosu**  
Special Telecommunications Service
- **Seema Shah**  
Vidyalankar Institute of Technology Mumbai
- **Seifedine Kadry**  
American University of the Middle East
- **Selem Charfi**  
HD Technology
- **SENGOTTUVELAN P**  
Anna University, Chennai
- **Senol Piskin**  
Istanbul Technical University, Informatics Institute
- **Sérgio Ferreira**  
School of Education and Psychology, Portuguese  
Catholic University
- **Seyed Hamidreza Mohades Kasaei**  
University of Isfahan
- **Shafiqul Abidin**  
HMR Institute of Technology & Management  
(Affiliated to GGSIP University), Hamidpur, Delhi -  
110036
- **Shahanawaj Ahamad**  
The University of Al-Kharj
- **Shaidah Jusoh**
- **Shaiful Bakri Ismail**
- **Shakir Khan**  
Al-Imam Muhammad Ibn Saud Islamic University
- **Shawki Al-Dubae**  
Assistant Professor
- **Sherif Hussein**  
Mansoura University
- **Shriram Vasudevan**  
Amrita University
- **Siddhartha Jonnalagadda**  
Mayo Clinic
- **Sim-Hui Tee**  
Multimedia University
- **Simon Ewedafe**  
The University of the West Indies
- **Siniša Opic**  
University of Zagreb, Faculty of Teacher Education
- **Sivakumar Poruran**  
SKP ENGINEERING COLLEGE
- **Slim BEN SAOUD**  
National Institute of Applied Sciences and  
Technology
- **Sofien Mhatli**
- **sofyan Hayajneh**
- **Sohail Jabbar**  
Bahria University
- **Sri Devi Ravana**  
University of Malaya
- **Sudarson Jena**  
GITAM University, Hyderabad
- **Suhail Sami Owais Owais**
- **Suhas J Manangi**  
Microsoft
- **SUKUMAR SENTHILKUMAR**  
Universiti Sains Malaysia
- **Süleyman Eken**  
Kocaeli University
- **Sumazly Sulaiman**  
Institute of Space Science (ANGKASA), Universiti  
Kebangsaan Malaysia

- **Sumit Goyal**  
National Dairy Research Institute
  - **Supareerk Janjarasjitt**  
Ubon Ratchathani University
  - **Suresh Sankaranarayanan**  
Institut Teknologi Brunei
  - **Susarla Sastry**  
JNTUK, Kakinada
  - **Suseendran G**  
Vels University, Chennai
  - **Suxing Liu**  
Arkansas State University
  - **Syed Ali**  
SMI University Karachi Pakistan
  - **T C.Manjunath**  
HKBK College of Engg
  - **T V Narayana rao Rao**  
SNIST
  - **T. V. Prasad**  
Lingaya's University
  - **Taiwo Ayodele**  
Infonetmedia/University of Portsmouth
  - **Talal Bonny**  
Department of Electrical and Computer Engineering, Sharjah University, UAE
  - **Tamara Zhukabayeva**
  - **Tarek Gharib**  
Ain Shams University
  - **thabet slimani**  
College of Computer Science and Information Technology
  - **Totok Biyanto**  
Engineering Physics, ITS Surabaya
  - **Touati Youcef**  
Computer sce Lab LIASD - University of Paris 8
  - **Tran Sang**  
IT Faculty - Vinh University - Vietnam
  - **Tsvetanka Georgieva-Trifonova**  
University of Veliko Tarnovo
  - **Uchechukwu Awada**  
Dalian University of Technology
  - **Udai Pratap Rao**
  - **Urmila Shrawankar**  
GHRCE, Nagpur, India
  - **Vaka MOHAN**  
TRR COLLEGE OF ENGINEERING
  - **VENKATESH JAGANATHAN**
- ANNA UNIVERSITY
  - **Vinayak Bairagi**  
AISSMS Institute of Information Technology, Pune
  - **Vishnu Mishra**  
SVNIT, Surat
  - **Vitus Lam**  
The University of Hong Kong
  - **VUDA SREENIVASARAO**  
PROFESSOR AND DEAN, St.Mary's Integrated Campus, Hyderabad
  - **Wali Mashwani**  
Kohat University of Science & Technology (KUST)
  - **Wei Wei**  
Xi'an Univ. of Tech.
  - **Wenbin Chen**  
360Fly
  - **Xi Zhang**  
illinois Institute of Technology
  - **Xiaojing Xiang**  
AT&T Labs
  - **Xiaolong Wang**  
University of Delaware
  - **Yanping Huang**
  - **Yao-Chin Wang**
  - **Yasser Albagory**  
College of Computers and Information Technology, Taif University, Saudi Arabia
  - **Yasser Alginahi**
  - **Yi Fei Wang**  
The University of British Columbia
  - **Yihong Yuan**  
University of California Santa Barbara
  - **Yilun Shang**  
Tongji University
  - **Yu Qi**  
Mesh Capital LLC
  - **Zacchaeus Omogbadegun**  
Covenant University
  - **Zairi Rizman**  
Universiti Teknologi MARA
  - **Zarul Zaaba**  
Universiti Sains Malaysia
  - **Zenzo Ncube**  
North West University
  - **Zhao Zhang**  
Deptment of EE, City University of Hong Kong
  - **Zhihan Lv**



Chinese Academy of Science

- **Zhixin Chen**  
ILX Lightwave Corporation
- **Ziyue Xu**  
National Institutes of Health, Bethesda, MD

- **Zlatko Stapic**  
University of Zagreb, Faculty of Organization and  
Informatics Varazdin
- **Zuraini Ismail**  
Universiti Teknologi Malaysia

# CONTENTS

Paper 1: A Robust Algorithm of Forgery Detection in Copy-Move and Spliced Images

Authors: Tu Huynh-Kha, Thuong Le-Tien, Synh Ha-Viet-Uyen, Khoa Huynh-Van, Marie Luong

PAGE 1 – 8

Paper 2: Portable Facial Recognition Jukebox Using Fisherfaces (Frj)

Authors: Richard Mo, Adnan Shaout

PAGE 9 – 14

Paper 3: Internet of Everything (Ioe): Analysing the Individual Concerns Over Privacy Enhancing Technologies (Pets)

Authors: Asim Majeed, Rehan Bhana, Anwar Ul Haq, Imani Kyaruzi, Shaheed Pervaz, Mike-Lloyd Williams

PAGE 15 – 22

Paper 4: A General Evaluation Framework for Text Based Conversational Agent

Authors: Mohammad Hijjawi, Zuhair Bandar, Keeley Crockett

PAGE 23 – 33

Paper 5: Effects of Walls and Floors in Indoor Localization Using Tracking Algorithm

Authors: Farhat M. A. Zargoun, Ibrahim M. Henawy, Nesreen I. Ziedan

PAGE 34 – 39

Paper 6: E-Learning Collaborative System for Practicing Foreign Languages with Native Speakers

Authors: Ilya V. Osipov, Alex A. Volinsky, Anna Y. Prasikova

PAGE 40 – 45

Paper 7: Color Image Segmentation via Improved K-Means Algorithm

Authors: Ajay Kumar, Shishir Kumar

PAGE 46 – 53

Paper 8: Detection and Feature Extraction of Collective Activity in Human-Computer Interaction

Authors: Ioannis Karydis, Markos Avlonitis, Phivos Mylonas, Spyros Sioutas

PAGE 54 – 59

Paper 9: NMVSA Greedy Solution for Vertex Cover Problem

Authors: Mohammed Eshtay, Azzam Sliet, Ahmad Sharieh

PAGE 60 – 64

Paper 10: Implementation of Pedestrian Dynamic

Authors: Purba Daru Kusuma

PAGE 65 – 70

Paper 11: Critical Path Reduction of Distributed Arithmetic Based FIR Filter

Authors: Sunita Badave, Anjali Bhalchandra

PAGE 71 – 78

Paper 12: Explorative Study of SQL Injection Attacks and Mechanisms to Secure Web Application Database- A Review

Authors: Chandershekhar Sharma, Dr. S. C. Jain, Dr. Arvind K Sharma

PAGE 79 – 87

**Paper 13: Corrupted MP4 Carving Using MP4-Karver**

*Authors: Ahmed Nur Elmi Abdi, Kamaruddin Malik Mohamad, Yusoof Mohammed Hasheem, Rashid Naseem, Jamaluddin, Muhammad Aamir*

**PAGE 88 – 93**

**Paper 14: Competitive Representation Based Classification Using Facial Noise Detection**

*Authors: Tao Liu, Cong Li, Ying Liu, Chao Li*

**PAGE 94 – 99**

**Paper 15: Energy Efficient Routing Protocol for Maximizing the Lifetime in Wsns Using Ant Colony Algorithm and Artificial Immune System**

*Authors: Safaa Khudair Leabi, Turki Younis Abdalla*

**PAGE 100 – 108**

**Paper 16: Improving DNA Computing Using Evolutionary Techniques**

*Authors: Godar J. Ibrahim, Tarik A. Rashid, Ahmed T. Sadiq*

**PAGE 109 – 121**

**Paper 17: A Novel Paradigm for Symmetric Cryptosystem**

*Authors: Shadi R. Masadeh, Hamza A. Al\_Sewadi, Mohammad A. Wadi*

**PAGE 122 – 127**

**Paper 18: Integrating Semantic Features for Enhancing Arabic Named Entity Recognition**

*Authors: Hamzah A. Alsayadi, Abeer M. ElKorany*

**PAGE 128 – 136**

**Paper 19: Clustering Analysis of Wireless Sensor Network Based on Network Coding with Low-Density Parity Check**

*Authors: Maria Hammouti, El Miloud Ar-reyouchi, Kamal Ghoumid, Ahmed Lichioui*

**PAGE 137 – 143**

**Paper 20: Multi-Agent Based Model for Web Service Composition**

*Authors: Karima Belmabrouk, Fatima Bendella, Maroua Bouzid*

**PAGE 144 – 150**

**Paper 21: A Novel Mapreduce Lift Association Rule Mining Algorithm (Mrlar) for Big Data**

*Authors: Nour E. Oweis, Mohamed Mostafa Fouad, Sami R. Oweis, Suhail S. Owais, Vaclav Snasel*

**PAGE 151 – 157**

**Paper 22: Knowledge Management of Best Practices in a Collaborative Environment**

*Authors: Amal Al-Rasheed, Jawad Berri*

**PAGE 158 – 165**

**Paper 23: An Automated Recommender System for Course Selection**

*Authors: Amer Al-Badarenah, Jamal Alsakran*

**PAGE 166 – 175**

**Paper 24: Fault Tolerant System for Sparse Traffic Grooming in Optical WDM Mesh Networks Using Combiner Queue**

*Authors: Sandip R. Shinde, Dr. Suhas H. Patil, Dr. S. Emalda Roslin, Archana S. Shinde*

**PAGE 176 – 180**



Paper 25: The ECG Signal Compression Using an Efficient Algorithm Based on the DWT

*Authors: Oussama El B'charri, Rachid Latif, Wissam Jenkal, Abdenbi Abenaou*

**PAGE 181 – 187**

Paper 26: Amba Based Advanced DMA Controller for SoC

*Authors: Abdullah Aljumah, Mohammed Altaf Ahmed*

**PAGE 188 – 193**

Paper 27: Role Based Multi-Agent System for E-Learning (MASel)

*Authors: Mustafa Hameed, Nadeem Akhtar, Malik Saad Missen*

**PAGE 194 – 200**

Paper 28: Detection and Identification System of Bacteria and Bacterial Endotoxin Based on Raman Spectroscopy

*Authors: Muhammad Elsayeh, Ahmed H.Kandil*

**PAGE 201 – 205**

Paper 29: Feature Based Correspondence: A Comparative Study on Image Matching Algorithms

*Authors: Usman Muhammad Babri, Munim Tanvir, Khurram Khurshid*

**PAGE 206 – 210**

Paper 30: New Mathematical Modeling of Three-Level Supply Chain with Multiple Transportation Vehicles and Different Manufacturers

*Authors: Amir Sadeghi, Amir Farmahini Farahani, Hossein Beiki*

**PAGE 211 – 217**

Paper 31: Wiki-Based Stochastic Programming and Statistical Modeling System for the Cloud

*Authors: Vaidas Giedrimas, Leonidas Sakalauskas, Marius Neimantas, Kestutis Žilinskas, Nerijus Barauskas, Remigijus Valčiukas*

**PAGE 218 – 223**

Paper 32: Multi Agent Architecture for Search Engine

*Authors: Disha Verma, Dr. Barjesh Kochar*

**PAGE 224 – 229**

Paper 33: Towards a New Approach to Improve the Classification Accuracy of the Kohonen's Self-Organizing Map During Learning Process

*Authors: El Khatir HAIMOUDI, Hanane FAKHOURI, Loubna CHERRAT, Mostafa Ezziyyani*

**PAGE 230 – 236**

Paper 34: A New Approach for Time Series Forecasting: Bayesian Enhanced by Fractional Brownian Motion with Application to Rainfall Series

*Authors: Cristian Rodriguez Rivero, Daniel Patiño, Julian Pucheta, Victor Sauchelli*

**PAGE 237 – 244**

Paper 35: Performance Evaluation of Content Based Image Retrieval on Feature Optimization and Selection Using Swarm Intelligence

*Authors: Kirti Jain, Dr.Sarita Singh Bhadauria*

**PAGE 245 – 249**

Paper 36: Application of Artificial Neural Networks for Predicting Generated Wind Power

*Authors: Vijendra Singh*

**PAGE 250 – 253**

Paper 37: Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data

*Authors: Suhail Sami Owais, Nada Sael Hussein*

**PAGE 254 – 258**

Paper 38: Hybrid Solution Methodology: Heuristic-Metaheuristic-Implicit Enumeration 1-0 for the Capacitated Vehicle Routing Problem (Cvrp)

*Authors: David Escobar Vargas, Ramón A. Gallego Rendón, Antonio Escobar Zuluaga*

**PAGE 259 – 268**

Paper 39: Automation of Optimized Gabor Filter Parameter Selection for Road Cracks Detection

*Authors: Haris Ahmad Khan, M. Salman, Sajid Hussain, Khurram Khurshid*

**PAGE 269 – 275**

Paper 40: Classification of Hand Gestures Using Gabor Filter with Bayesian and Naïve Bayes Classifier

*Authors: Tahira Ashfaq, Khurram Khurshid*

**PAGE 276 – 279**

Paper 41: Moon Landing Trajectory Optimization

*Authors: Ibrahim Mustafa MEHEDI, Md. Shofiqui ISLAM*

**PAGE 280 – 286**

Paper 42: Evaluation of Navigational Aspects of Moodle

*Authors: Raheela Arshad, Awais Majeed, Hammad Afzal, Muhammad Muzammal, Arif ur Rahman*

**PAGE 287 – 298**

Paper 43: Communication-Load Impact on the Performance of Processor Allocation Strategies in 2-D Mesh Multicomputer Systems

*Authors: Zaid Mustafa, J. J. Alshaer, O. Dorgham, S. Bani-Ahmad*

**PAGE 299 – 310**

Paper 44: ECG Signal Compression Using the High Frequency Components of Wavelet Transform

*Authors: Surekha K.S, B. P. Patil*

**PAGE 311 – 315**

Paper 45: Feature Selection Based on Minimum Overlap Probability (MOP) in Identifying Beef and Pork

*Authors: Khoerul Anwar, Agus Harjoko, Suharto Suharto*

**PAGE 316 – 322**

Paper 46: Resource Allocation in Cloud Computing Using Imperialist Competitive Algorithm with Reliability Approach

*Authors: Maryam Fayazi, Mohammad Reza Noorimehr, Sayed Enayatollah Alavi*

**PAGE 323 – 331**

Paper 47: Real-Time Gender Classification by Face

*Authors: Eman Fares Al Mashagba*

**PAGE 332 – 336**

Paper 48: An Accelerated Architecture Based on GPU and Multi-Processor Design for Fingerprint Recognition

*Authors: Mossaad Ben Ayed, Sabeur Elkosantini*

**PAGE 337 – 348**

**Paper 49: Planning And Allocation of Tasks in a Multiprocessor System as a Multi-Objective Problem and its Resolution Using Evolutionary Programming**

*Authors: Apolinar Velarde Martinez, Eunice Ponce de León Sentí, Juan Antonio Nungaray Ornelas, Juan Alejandro Montañez de la Torre*

**PAGE 349 – 360**

**Paper 50: An Improved Image Steganography Method Based on LSB Technique with Random Pixel Selection**

*Authors: Marwa M. Emam, Abdelmgeid A. Aly, Fatma A. Omara*

**PAGE 361 – 366**

**Paper 51: Developing a Feasible and Maintainable Ontology for Automatic Landscape Design**

*Authors: Pintescu Alina, Matei Oliviu-Dorin, Boancă Iuliana Păunița, Honoriu Vălean*

**PAGE 367 – 374**

**Paper 52: Parallel Implementation of Bias Field Correction Fuzzy C-Means Algorithm for Image Segmentation**

*Authors: Nouredine AITALI, Bouchaib CHERRADI, Ahmed EL ABBASSI, Omar BOUATTANE, Mohamed YOUSSEFI*

**PAGE 375 – 383**

**Paper 53: An Algerian dialect: Study and Resources**

*Authors: Salima Harrat , Karima Meftouhy, Mourad Abbasz, Khaled-Walid Hidoucix, Kamel Smaili*

**PAGE 384 – 396**

**Paper 54: An Enhanced Automated Test Item Creation Based on Learners Preferred Concept Space**

*Authors: Mohammad AL-Smadi, Margit Hoffler, Christian Groll*

**PAGE 397 – 405**

**Paper 55: Characterizing End-to-End Delay Performance of Randomized TCP Using an Analytical Model**

*Authors: Mohammad Shorfuzzaman, Mehedi Masud, Md. Mahfuzur Rahman*

**PAGE 406 – 412**

**Paper 56: Improving Vertical Handoffs Using Mobility Prediction**

*Authors: Mahmoud Al-Ayyoub, Ghaiith Husari, Wail Mardini*

**PAGE 413 – 419**

**Paper 57: Performance Evaluation of Affinity Propagation Approaches on Data Clustering**

*Authors: R. Refianti, A.B. Mutiara, A.A. Syamsudduha*

**PAGE 420 – 429**

**Paper 58: Modified Grapheme Encoding and Phonemic Rule to Improve PNNR-Based Indonesian G2P**

*Authors: Suyanto, Sri Hartati, Agus Harjoko*

**PAGE 430 – 435**

**Paper 59: Testing and Analysis of Activities of Daily Living Data with Machine Learning Algorithms**

*Authors: Ayse Cufoglu, Adem Coskun*

**PAGE 436 – 441**

**Paper 60: A Context-Aware Recommender System for Personalized Places in Mobile Applications**

*Authors: Soha A.El-Moemen Mohamed, Taysir Hassan A.Soliman, Adel A.Sewis*

**PAGE 442 – 448**

**Paper 61: Performance Enhancement of Patch-based Descriptors for Image Copy Detection**

*Authors: Junaid Baber, Maheen Bakhtyar, Waheed Noor, Abdul Basit, Ihsan Ullah*

**PAGE 449 – 456**

# A Robust Algorithm of Forgery Detection in Copy-Move and Spliced Images

Tu Huynh-Kha<sup>1,2</sup>, Thuong Le-Tien<sup>1</sup>, Synh Ha-Viet-Uyen<sup>1</sup>, Khoa Huynh-Van<sup>1</sup>, Marie Luong<sup>3</sup>

<sup>1</sup>Department of Electronics and Electrical Engineering, Ho Chi Minh City University of Technology, Vietnam

<sup>2</sup>School of Computer Science and Engineering, International University-VNUHCM, Vietnam

<sup>3</sup>Laboratory L2TI, University Paris 13, France

**Abstract**—The paper presents a new method to detect forgery by copy-move, splicing or both in the same image. Multiscale, which limits the computational complexity, is used to check if there is any counterfeit in the image. By applying one-level Discrete Wavelet Transform, the sharpened edges, which are traces of cut-paste manipulation, are high frequencies and detected from LH, HL and HH sub-bands. A threshold is proposed to filter the suspicious edges and the morphological operation is applied to reconstruct the boundaries of forged regions. If there is no shape produced by dilation or no highlight sharpened edges, the image is not faked. In case of forgery image, if a region at the other position is similar to the defined region in the image, a copy-move is confirmed. If not, a splicing is detected. The suspicious region is extracted the feature using Run Difference Method (RDM) and a feature vector is created. Searching regions having the same feature vector is called detection phase. The algorithm applying multiscale and morphological operation to detect the sharpened edges and RDM to extract the image features is simulated in Matlab with high efficiency not only in the copy-move or spliced images but also the image with both copy-move and splicing.

**Keywords**—Forgery detection (FD); Copy-Move; Discrete Wavelet Transform (DWT); Run Difference Method (RDM); Splicing, Sharpness

## I. INTRODUCTION

Image forgery detection is attracting the attention of scientists in computer vision, digital image processing, biomedical technology, investigation, forensics, etc. With popular and complicated technologies and powerful software tools in digital images, it is difficult to confirm if the image is original or not through naked eyes (see Figure 1). This challenges researchers to develop algorithms and propose methods to detect the forgery in image. Upon the survey on IEEE and Elsevier, the number of publications on image forgery detection from 2000 increases rapidly in 2010 and more in the following years [1]. An image can be faked by changing any characteristics including brightness, darkness or image parameters,... or hiding information. Watermarking and digital signature are solutions of information security in which a security code is inserted in the image so these methods have information of a code and the original image. A question is asked in the case if there is no code or signature inserted or information of original image, how to confirm its authenticity. Blind/passive techniques in which the detection is done in the tested image itself without any prior information are developed to solve the problem given.

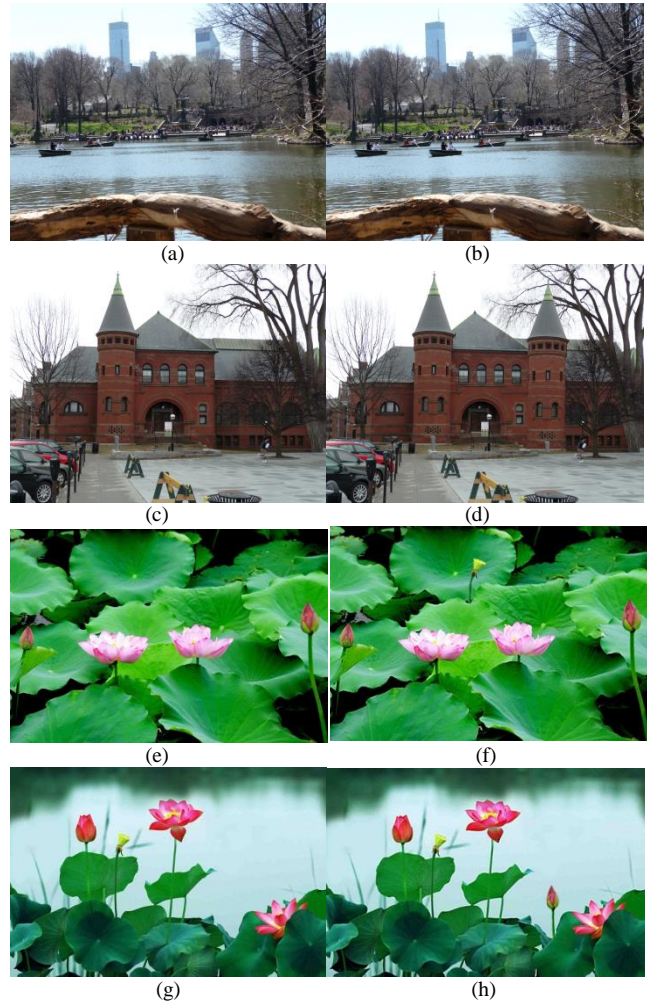


Fig. 1. Original, Copy-Move and Spliced Images in [2] and by Photoshop.

(a), (c), (e), (g). Original images.

(b), (d). Copy-move images from (a), (c) from [2].

(f), (h). Spliced images from (e), (g) by Photoshop.

According to [3], blind/passive techniques are grouped into two kinds: copy-move and splicing. The copy-move is defined by cutting an image region and pasting it to other place in the same image while splicing is understood by cutting an image region and pasting it to a different image. Based on this classification, searching the regions having similar features in copy-move images or completely different regions in spliced images is the principle of forgery detection.

Many techniques are proposed and used in this field but actually, they can solve only problems on copy-move or splicing separately. The dataset in the previous publications often consists of copy-move images or spliced images, not both in images. This paper proposes a method which can detect the forgery in images not only for copy-move or splicing but also for both. The literature review and proposed method are presented in part II and III. Simulation results and conclusion are shown in the following parts.

## II. LITERATURE REVIEW

This part summarizes some recent methods related to image forgery detection as an overview and also references from which the researchers have new ideas and solutions. For copy-move detection, the searching of similar regions is the main purpose in almost all methods while the searching of inconsistencies of features is considered the solution in splicing detection. Although there are numerous related methods published, most of them solve problems of copy-move or splicing separately and only few papers can solve problems of both copy-move and splicing in the same image. Therefore, developing an algorithm to detect any forgery regions, not limited to copy-move or forgery, is still a challenge for scientists in the field of image forensics.

### A. Copy-Move Forgery Detection

For copy-move detection, a survey in [3] covers and evaluates methods published until 2012 in which the duplicated regions are confirmed based on feature vectors comparisons. Feature vectors can be extracted directly from tested images or after applying a transformation such as DWT and DCT. The difference on feature extractions and the way to compare feature vectors comprise the variety in methods. After that, a new method to extract the image features by describing the spatial structure of the gray image texture called Local Binary Pattern (LBP) was introduced by Leida Li *et al.* [4] in 2013. In the case of color image, it should be first converted to gray image by using  $I=0.299R+0.587G+0.114B$  and low pass filter should be applied to obtain the low-frequency features which is more stable than the high-frequency ones. As the previous methods, the feature matching is defined based on the threshold. Moreover, the post-processing including a special designed filter and morphological operations is also considered in the process of detection. The method is robust to JPEG compression, noise contamination, blurring, rotation and flipping. However, it is difficult to detect the rotated regions with general angles. Investigation of invariant block features and appropriate selection of the dimension of features are suggested to improve the random rotation.

Using Undecimated Dyadic Wavelet Transform (UDWT) and Zernike moments is proposed as a new method to detect the forgery in copy-move images by Jiyun Yang [5] in 2013. In this paper, the applying UDWT is firstly used to collect the low frequencies (LL) components. Traditional ZMs is then computed to produce feature vectors of overlapping blocks on LL and find the copied regions from these vectors. Lexicographical sorting, correlation coefficients with a threshold value are used to find the similar vectors and limit the exact forged blocks from the groups of similar vectors obtained in lexicographical sorting step, respectively.

Blur invariants are also used to produce feature vectors in copy-move image forgery detection [6]. Based on this idea, the LL sub-band from DWT of an image using Harr basis is divided into small overlapping blocks whose features are then represented by blur moment invariants. Each block feature vector consists of 24 blur invariants in case of grayscale images and 72 ones in the RGB and is reduced dimension by applying PCA. The block similarity analysis will detect the duplicated regions by considering the Euclidean distances and a user-defined threshold. This is applied to image with noise, blur and contrast changes. The applying other basis or DCT is suggested in the coming research.

In [7], an image is decomposed into four sub-bands using DWT in which the LL sub-band is considered for the coming steps. The proposed algorithm uses SIFT on each small overlapping block divided from LL sub-band to extract feature vector. These feature vectors are used to create a descriptor vector and compared to detect if there is a copy-move manipulation in the image. This method is checked with MICC-F200 database with high accuracy, less time, robust to scale and rotation. The authors of this algorithm developed SUFT (Speed-Up Robust Feature) to extract the features of image block instead of using SIFT as in [7]. The combination of SUFT and DWT, DyWT are also presented. With the results obtained from the proposed method, SUFT is proved faster than SIFT while SIFT is mostly used to select the invariant features [8].

### B. Splicing Forgery Detection

Splicing is more complex than copy-move, not only in the forgery manipulation but also in detection. The key idea of many splicing detection methods is searching regions being inconsistent with camera characteristics or image features. Regions which are resampled, double compressed, and those with blur discrepancies or sharpness differences can be considered traces of splicing. However, because of the variety of splicing, more and more algorithms have been developed in recent years.

Conditional Co-occurrence Probability Matrix (CCPM) is used to detect the splicing in image based on the third order statistical features [9]. CCPM contains the discriminative information which are included in higher order statistical features and independent to the image features. However, the higher dimensionality of features is, the more complex computation is. Therefore, Principle Component Analysis (PCA) is also used to improve the computational complexity of the proposed method which is robust and better than Markov features both in spatial domain and block discrete cosine transform (BDCT) domain.

Rescaling and its factor are used to detect the forgery caused by splicing [10]. A region copied from an image will be resized or scaled before pasting to the destination image. Scaling makes the pasted portion resampled and inconsistent. In addition, properties of the zero-crossing of the second difference are considered to calculate the scaling factor with different interpolation schemes. The algorithm of rescale detection and estimation was proposed clearly in five steps including pre-processing to convert the RGB to grayscale and extract Y component from  $YCbCr$  conversion; calculate the



second difference, their zero-crossing and Discrete Fourier Transform (DFT) before searching for the periodicity and peak detection.

Differences of JPEG compression in an image can be caused by the splicing [11]. JPEG forgery detection based on 8x8 block Discrete Cosine Transform (DCT) transform to detect the shift of DCT block alignment. The splicing detection was proposed by analyzing and suggesting solutions for cases making the differences in compression history including detections of Aligned Double JPEG, Non Aligned Double JPEG, Primary Quantization Table, JPEG ghost.

Illumination inconsistencies and intrinsic resampling properties are also parameters to detect the splicing [12]. The first requires an input image and a database for training. The algorithm begins with 30x30 blocks which will be transformed into an opponent color space HSV before extracting features of contrast and mean. The contrast is calculated from the standard deviation while the mean is obtained by computing the average grey level. These features will decide suitable algorithm. Illumination color estimation, illumination map creation, Wavelet-based features extraction and classifier are the following steps of the proposed method by illumination differences detection. The second solution in this paper proposes a resampling detection scheme to detect forgery in which second difference in horizontal or vertical, Radon transform, FFT of covariance, high-pass filtering, feature extraction and classifier are included.

### C. Forgery Detection for both Copy-Move and Splicing

An integrated technique, which combines DCT and Speeded Up Robust Features (SURF), to detect the image forgery in term of copy-move or splicing was proposed in 2011 [13]. This means the tested images can be optional, not classified in copy-move or splicing in advance. The paper finds new traces based on recompression to detect the counterfeit of recompressed images. Periodicity analysis with double compression effect in both spatial and DCT domain is applied before using SURF descriptor to against the variation of rotation and scaling. The proposed method located the forgery regions efficiently for both copy-move and splicing image, especially, discriminated the positions of original and forged regions.

At the EUROCON 2013 in Croatia, a method can detect both copy-move and splicing in image using a multi-resolution Web Law Descriptor (WLD) was presented [14]. The algorithm firstly converts a RGB image into  $YCbCr$  so that the WLD can extract the features from chrominance components which are less sensitive than luminance. To extract the features effectively, WDM are expressed by two components of differential excitation and orientation, which based on Weber's law. The multi-resolution WLD histogram is comprised of the histogram of three neighbors of (8,1), (16,2), and (24,3) where the first argument is the number of neighboring pixels and the second is the radius of the neighbors from the center pixel. A support vector machine, which involves to training and testing image, is used for classification purpose.

## III. PROPOSED METHOD

The paper consists of two phases: the sharpened edge detection and the copy-move/splicing detection. Before presenting the proposed method, this part firstly shows the related theories in brief including multiscale using DWT, edge detection, dilation and RDM in which the first three [15] are used in sharpness detection and the last is suggested for feature extraction in copy-move/splicing detection.

### A. Multiscale using DWT

With a 2D image  $f(x,y)$ , two dimension DWT will produce one separable scaling function  $\phi(x,y)$  and three separable directionally sensitive wavelets  $\psi^H(x,y)$ ,  $\psi^V(x,y)$ ,  $\psi^D(x,y)$  corresponding to variations along the horizontal edges, vertical edges and diagonals, respectively. These functions are defined in (1), (2), (3) and (4).

$$\phi(x,y) = \phi(x)\phi(y) \quad (1)$$

$$\psi^H(x,y) = \psi(x)\phi(y) \quad (2)$$

$$\psi^V(x,y) = \phi(x)\psi(y) \quad (3)$$

$$\psi^D(x,y) = \psi(x)\psi(y) \quad (4)$$

where  $\phi(x)$ ,  $\phi(y)$  are one dimension scaling functions and  $\psi(x)$ ,  $\psi(y)$  are one dimension wavelet functions.

In DWT, a scaling function is used to create a series of approximation of an image and a factor of 2 in resolution defines the difference between its nearest neighboring approximations while the encoding of differences in information between adjacent approximations is obtained from wavelets. The scaled and translated basic functions are defined by (5) and (6)

$$\phi_{j,m,n}(x,y) = 2^{j/2} \phi(2^j x - m, 2^j y - n) \quad (5)$$

$$\psi_{j,m,n}^i(x,y) = 2^{j/2} \psi^i(2^j x - m, 2^j y - n) \quad (6)$$

for all  $j, k \in \mathbf{Z}$ ,  $m=n=0,1,2,\dots,2^j-1$ . In (6),  $i=\{H,V,D\}$  identifies the directional wavelets from (2), (3) and (4). Then discrete wavelet transform of image  $f(x,y)$  of size  $M \times N$  is done by defining the approximation and directional coefficients as in (7) and (8).

$$W_\phi(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \phi_{j_0,m,n}(x,y) \quad (7)$$

$$W_\psi^i(j, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \psi_{j,m,n}^i(x,y) \quad (8)$$

where  $j_0$  is an arbitrary scale,  $W_\phi(j_0,m,n)$  are approximation coefficients of image  $f(x,y)$  at scale  $j_0$  and  $W_\psi^i(j,m,n)$  are coefficients used to add the horizontal, vertical and diagonal details for scale  $j \geq j_0$ .

After applying DWT, an image is decomposed in approximation, horizontal, vertical and diagonal part (see Figure 2).

The edges are high frequencies which are collection of details in part II, III and IV of the Figure 2. As the proposed method reduces the size of image by a half so one-level DWT is applied. The filter bank to create one level-analysis is shown in Figure 3.

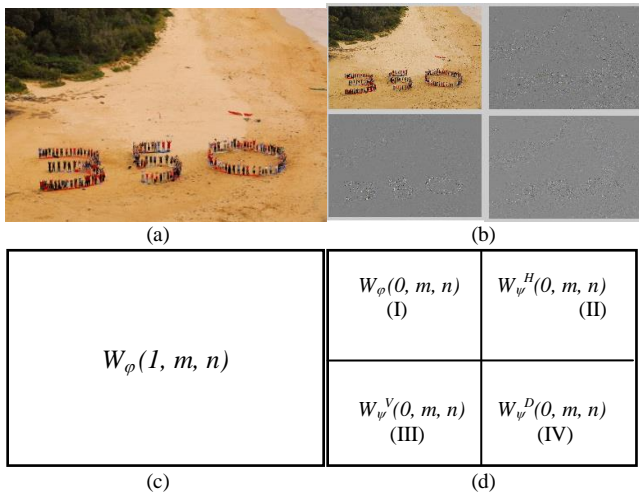


Fig. 2. One-level decomposition of a 2D image.

(a). The original image; (b). A one-level DWT.  
 (c), (d). Positions of corresponding sub-bands

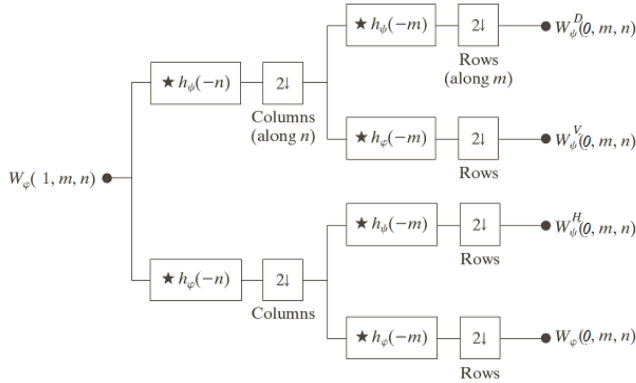


Fig. 3. One-level analysis filter bank

### B. Edge Detection

Sharpness of edges can be traces of pasting information from other region. Therefore, edge detection is the first step to search the suspicious regions and the regions having edges with highest sharpness are collected, considered and tested [16]. Laplacian operator is applied to the three sub-bands LH, HL and HH to select only edges for further processing steps by a convolution between each sub-band and a 3x3 Laplace kernel (see Figure 4).

-1	-1	-1
-1	8	-1
-1	-1	-1

Fig. 4. A 3x3 Laplace kernel

0	1	0
1	1	1
0	1	0

Fig. 5. Structuring element for bridging gaps

### C. Dilation for Filling Gaps

Ordinary, at positions of pasting, the borders will be smoothed by some software tools or Photoshop so not all of edges are detected continuously in LH, HL and HH. Therefore,

dilation is proposed to bridge the gaps and make the boundary smooth, which helps to address the forged regions easier.

The dilation of two sets A and B in  $Z^2$  is defined in (9)

$$A \oplus B = \{z \mid (\hat{B})_z \cap A \neq \emptyset\} \quad (9)$$

or in another form as in (10)

$$A \oplus B = \{z \mid [(\hat{B})_z \cap A] \subseteq A\} \quad (10)$$

where  $\hat{B}$  is the reflection of B and  $(\hat{B})_z$  is the shifting of  $\hat{B}$  by z.

By applying each LH, HL and HH to A and the structuring element defined in Fig.5 to B, the dilation is applied to repair gaps in boundaries which are maybe traces of cutting and pasting.

### D. Extract Features using Run Difference Method [17]

Run Difference Method (RDM) is a features extraction method in which features of size and prominence of texture elements are considered. From distribution of gray-level difference (DGD), RDM calculates five feature vectors including large difference emphasis, sharpness, the second moment of DGD, the second moment of distribution of the average gray level difference (DOD) and long distance emphasis.

With a rectangular gray image F in domain D of 2-dimensional image plane, the relationship between F and D can be defined as in (11) and (12).

$$D = \{(x, y) : x, y \in I, x \in [0, N_x], y \in [0, N_y]\} \quad (11)$$

$$F = \{(x, y, k) : (x, y) \in D, k \in I, k \in [0, n_g], k = f(x, y)\} \quad (12)$$

where  $N_x$  and  $N_y$  are horizontal and vertical dimensions of F;  $n_g$  is number of gray levels in F and I is set of integers.

Let  $\bar{d}$  be the displacement vector between two pixels  $(x_1, y_1)$  and  $(x_2, y_2)$ , we have:

$$\bar{d} = \{[\Delta x, \Delta y] = [x_2, y_2] - [x_1, y_1] : (x_1, y_1), (x_2, y_2) \in D, \|\Delta x, \Delta y\| > 0\} \quad (13)$$

$\bar{d}$  can be presented with distance r and direction  $\theta$  in polar coordinate as in (14)

$$\bar{d}_\theta(r) = [\Delta x, \Delta y], r = \sqrt{\Delta x^2 + \Delta y^2}, \theta = \tan^{-1}\left(\frac{\Delta y}{\Delta x}\right) \quad (14)$$

The Run Difference Matrix is defined as a function of r and gray level difference with the given direction  $\theta$  in (15)

$$RDM(r, dif) = \#\{(x_1, y_1), (x_2, y_2) : (x_1, y_1), (x_2, y_2) \in D, [x_2, y_2] - [x_1, y_1] = \bar{d}_\theta(r) \mid f(x_1, y_1) - f(x_2, y_2) = dif\} / N \quad (15)$$

where # denotes the cardinality. The denominator N is a normalization factor which is equal to the total number of paired pixels.

$$N = \#\{(x_1, y_1), (x_2, y_2) : (x_1, y_1), (x_2, y_2) \in D, [x_2, y_2] - [x_1, y_1] = \bar{d}_\theta(r)\} \quad (16)$$

From Run Difference Matrix as in Figure 6, three vectors including Distribution of Gray Level Difference (DGD), Distribution of Average Difference (DOD) in each row of RDM and Distribution of Average Distance (DAD) are defined in (17), (18) and (19) respectively.

$$DGD(j) = \sum_{r=1}^c RDM(r, j) \quad (17)$$

where c is the maximum distance of r.

$$DOD(r) = \sum_{j=0}^{n_g} RDM(r, j).j \quad (18)$$

$$DAD(j) = \sum_{r=1}^c RDM(r, j).r \quad (19)$$

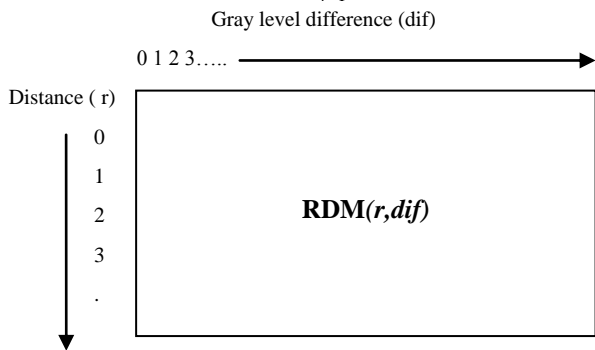


Fig. 6. Run Difference Matrix

Based on three vectors DGD, DOD, and DAD, five features including Large Difference Emphasis (LDE), Sharpness (SHP), Second Moment of DGD (SMG), Second Moment of DOD (SMO) and Long Distance Emphasis (LDEL) are also defined.

$$LDE = \sum_{j=0}^{n_g} DGD(j). \ln(K / j) \quad (20)$$

with K is a constant.

$$SHP = \sum_{j=0}^{n_g} DGD(j).j^3 \quad (21)$$

$$SMG = \sum_{j=0}^{n_g} DGD(j)^2 \quad (22)$$

$$SMO = \sum_{r=1}^c DOD(r)^2 \quad (23)$$

$$LDEL = \sum_{j=0}^{n_g} DAD(j).j^2 \quad (24)$$

These five parameters are considered the five features of images and used in image features extraction.

### E. Proposed Method

The paper proposes a method not only detecting the forgery but also defining the manipulation of forgery

including copy-move, splicing or both in an arbitrary image without any prior information of the original image. In addition, the method can detect more than one forged regions in an image. The flowchart of the proposed method, which consists of the edge detection to confirm forgery and similar region searching to define the forgery manipulation, is split and represented in Figure 7 and Figure 8.

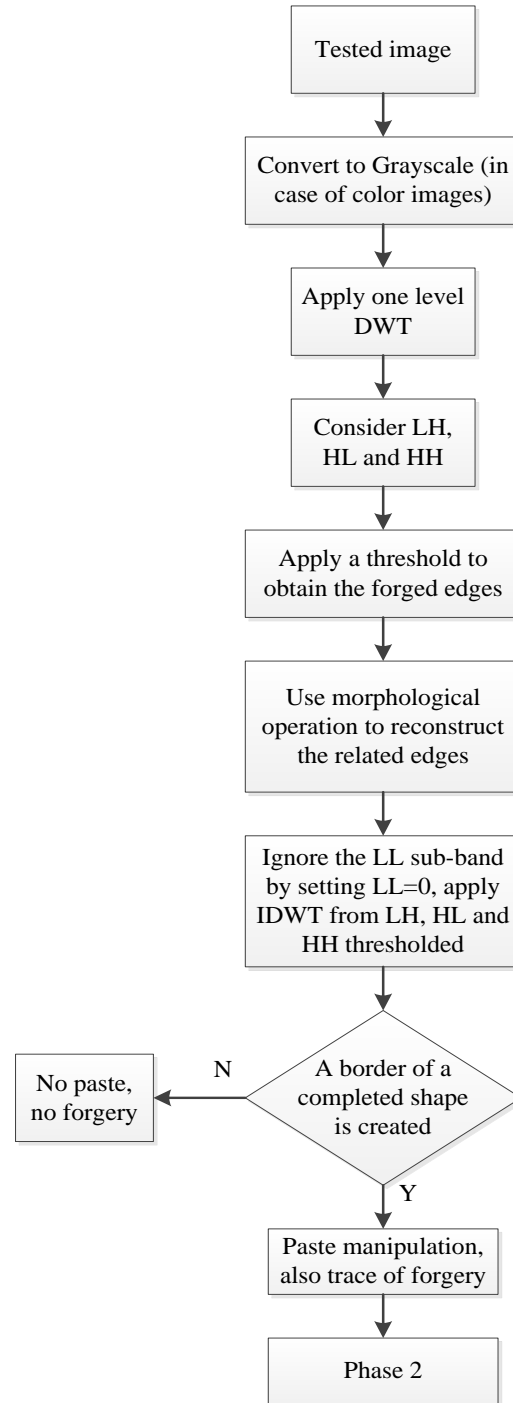


Fig. 7. Forgery confirmation based on boundary of the suspicious region

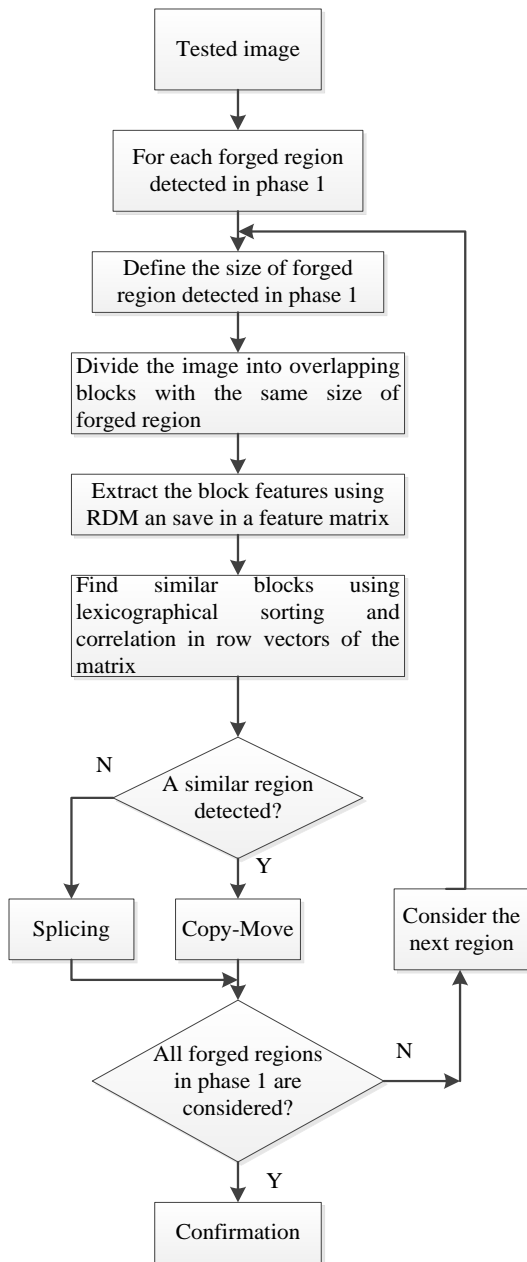


Fig. 8. The confirmation in kind of forgery: copy-move, splicing or both copy-move and splicing

In the first phase shown in Figure 7, a color image is converted to grayscale before applying one-level DWT decomposition. As edges are expressed by high frequencies, the three sub-bands LH, HL and HH are considered to detect the edges. Actually, there are many edges or boundaries in a real image so the collection of edges caused by pasting is required. The threshold is set up to the texture and layout of each image, and ranges from 50% to 80% of the maximum sharpness. The remaining edges after sharpening and filtering by thresholds in all three sub-bands of high frequencies are dilated to reconstruct edges or boundaries. To detect the cutting/pasting parts, the low frequencies in LL sub-band are ignored by setting them to zero. Therefore, Inverse discrete wavelet transform (IDWT) from these four sub-bands shows an

image with only edges and boundaries. If there is any feasible shape covered by edges, this is caused by a pasting so a counterfeit is confirmed. The number of completed shapes is the number of forged regions. Otherwise, the image is original.

For every faked part in Figure 8, copy-move or splicing manipulation is confirmed by feature similarity detection. Blob detection is applied to define the size  $m \times n$  of forged region. By dividing the tested  $M \times N$  image in many overlapping  $m \times n$  blocks,  $(M-m+1)(N-n+1)$  feature vectors are created by using Run Different Method. The algorithm detection uses Run Difference Method (RDM) extracts five features of the faked parts and searches regions having similar features. The results may be in these cases (i) copy-move if there is at least one other place having similar feature to the faked one, (ii) splicing if there is no similar region, (iii) both copy-move and splicing or more than one copy-move regions if there are at least two forged regions, the copy-move is defined as in (i) and the splicing is confirmed as in (ii).

#### IV. SIMULATION RESULT

The proposed algorithm is run in Matlab2013 by PC with processor Intel(R) Core™ i5-2400 CPU@3.10 GHz, RAM 4GB. The paper proposes an algorithm by using one-level DWT to address suspicious regions based on the edges with high sharpness from three sub-bands LH, HL and HH. The copy-move, splicing or both manipulations in an image are detected by searching regions similar to the suspicious regions. The test images for testing copy-move forgery are collected from the benchmark data of research group in [2]. The dataset for splicing and both copy-move and splicing are natural pictures and forged by Photoshop. Some results obtained from the proposed methods are shown in Figure 9.

#### Evaluation

The proposed method is evaluated based on three different data consisting of copy-move images, spliced images and both of copy-move and splicing in the same images. In the case of testing in copy-move images, the proposed algorithm is compared at image level to the Zernike moments [2], Undecimated Dyadic Wavelet Transform and Zernike Moments [18] and Discrete Wavelet Transform and Modified Zernike Moments [19] based on some images from benchmark\_data [2] with results as in Table 1.

Three parameters called precision ( $p$ ), recall ( $r$ ) and F1 are used to evaluate the feasibility of the proposed method, which defined in (25), (26) and (27) [2]. Precision is the probability of the exact forgery detection while recall is the probability of forged image detection. F1 is obtained by considering both precision and recall. To compare the efficiency between related methods, these parameters are calculated at image level and also in the set of copy-move images.

$$p = \frac{T_P}{T_P + F_P} \quad (25)$$

$$r = \frac{T_P}{T_P + F_N} \quad (26)$$

$$F_1 = 2 \cdot \frac{p \cdot r}{p + r} \quad (27)$$

where  $T_P$ ,  $F_P$  and  $F_N$  are the number of true forged pixels, false forged pixels and miss forged pixels, respectively.

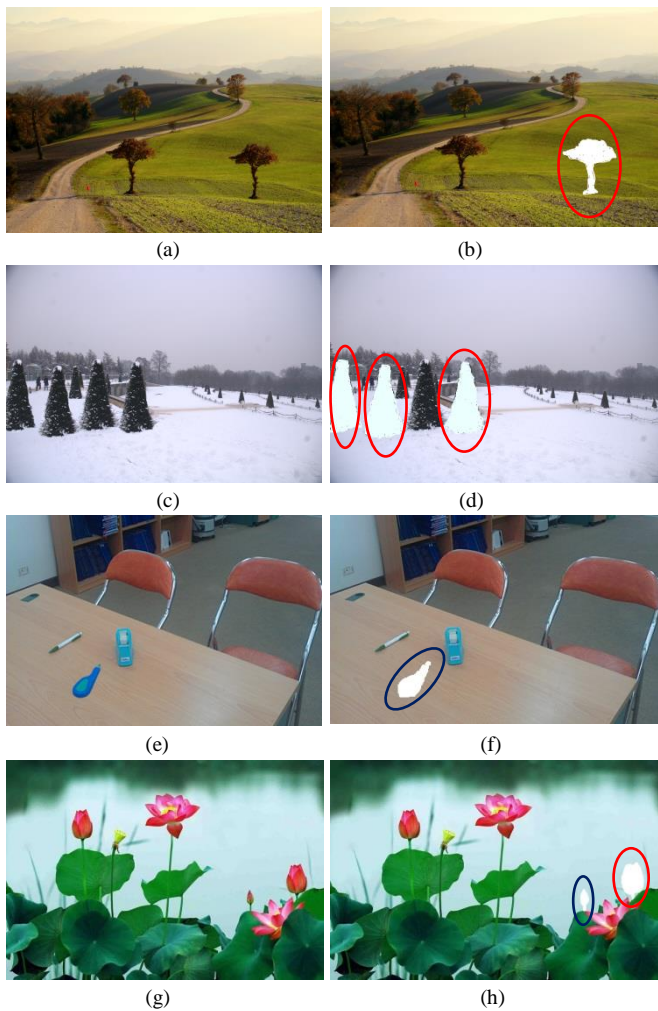


Fig. 9. Some simulation results by the proposed method.

- (a). Copy-move image from benchmark\_data [2]
  - (b). Copy-move detection of (a).
  - (c). Copy-move image from benchmark\_data [2]
  - (d). Copy-move detection of (c).
  - (e). Spliced image by Photoshop.
  - (f). Splicing detection of (e)
  - (g). Image with copy-move and splicing by Photoshop.
  - (h). Copy-move and splicing detection of (g).
- Red markers: copy-move; blue markers: splicing.

The efficiency of the proposed method for images with splicing or both copy-move and splicing, which are forged by Photoshop, are shown in Fig.9(f), Fig.9(h) and Fig.10.

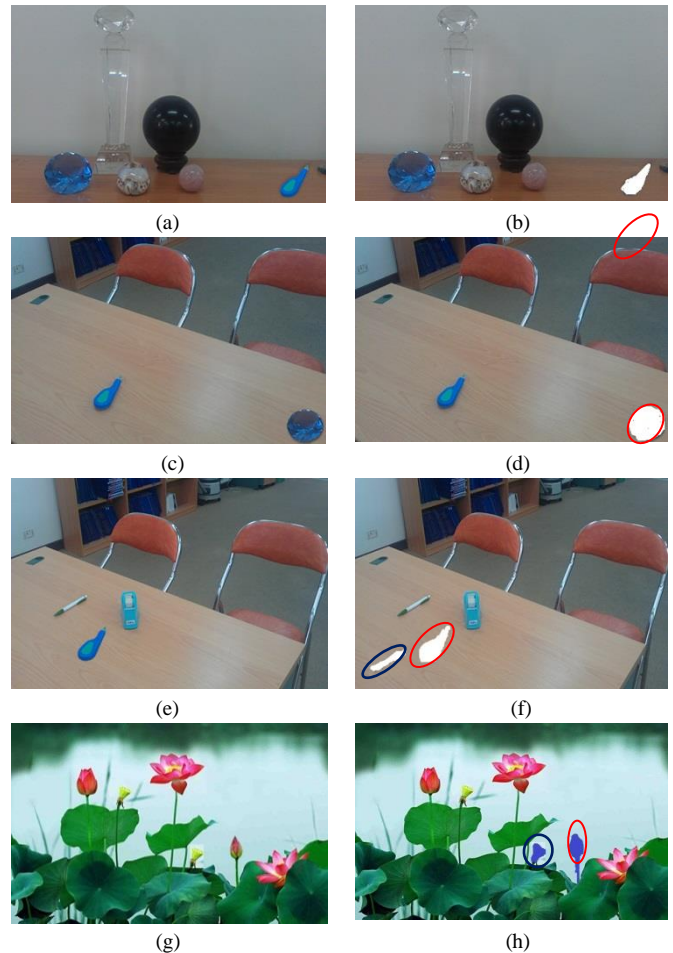


Fig. 10. Some simulation results in images, which are forged by splicing and both copy-move and splicing, obtained from the proposed method.

- (a), (c). Spliced images.
  - (b), (d). Splicing detection of (a) and (c), respectively.
  - (e),(g). Copy-move and spliced images.
  - (f), (h). Copy-Move and splicing detection of (e) and (g), respectively
- Red markers: copy-move; blue markers: splicing.

## V. CONCLUSION

The paper proposes a method to detect the forgery manipulations in images including copy-move or splicing or both. A counterfeit is firstly defined from the sharpness of edges and boundaries presented by high frequencies at the three sub-bands LH, HL and HH of one-level DWT decomposition which are traces of cutting and pasting. When a fake is confirmed, suspicious regions becomes objects to be considered. Through the blob detection, the size of suspicious parts are defined and the searching other places having similar RDM to these will classify the forgery of copy-move, splicing or both. The fact that tested images can be optional instead of limiting on copy-move or splicing is the novelty of proposed method. To evaluate the efficiency and feasibility of our

TABLE I. RESULTS FOR COPY-MOVE IMAGE DETECTION AT IMAGE LEVEL (%) IN CASE OF COPY-MOVE AT ONE PLACE AND FEW PLACES

Cases Methods	Copy-move detection (one place)			Copy-move detection (few places)		
	Precision (p)	Recall (r)	F <sub>1</sub>	Precision (p)	Recall (r)	F <sub>1</sub>
ZMs [2]	92.05	86.92	94.05	89.37	87.79	88.57
UDWT+ZMs [18]	91.87	85.32	94.05	90.24	86.05	88.1
DWT+MZMs [19]	92.68	88.16	94.05	91.09	87.07	89.03
Proposed method	94.05	89.27	94.05	93.79	90.25	91.99

method, the algorithm is tested in three different kinds of images in which the first kind is copy-move images from benchmark\_data [2] and two remains are spliced images and copy-move/spliced images by Photoshop with good results. Applying a Canny filter with suitable coefficients instead of using the morphological operation to limit the changes on energy of images can be considered in the coming research.

#### ACKNOWLEDGMENT

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2015-20-02.

#### REFERENCES

- [1] Gajanan K.B., Vijay H.M., "Digital image forgery detection using passive techniques: A survey", Digital Investigation, Elsevier Ltd. All rights reserved, April 2013.
- [2] V. Christlein, C. Riess, J. Jordan, C. Riess, E. Angelopoulou: "An Evaluation of Popular Copy-Move Forgery Detection Approaches", IEEE Transactions on Information Forensics and Security, Vol. 7, No. 6, pp. 1841-1854, 2012.
- [3] Tu H.K., Thuong L.T., Khoa H.V., Sy N.C., "A survey on Image Forgery Detection Techniques", The 11th IEEE-RIVF International Conference on Computing and Communication Technologies", Can Tho, Vietnam, Jan 25-28 2015.
- [4] Leida Li, Shushang Li, Hancheng Zhu, "An efficient scheme for detecting Copy-Move forged images by local binary patterns", Journal of Information Hiding and Multimedia Signal Processing, Vol. 4, No. 1, pp. 46-56, January 2013.
- [5] Jiyun Yang, Pei Ran, Di Xiao, Jinyong Tan, "Digital Image Forgery Forensics by Undecimated Dyadic Wavelet Transform and Zernike Moments", Journal of Computational Information Systems, No.9, pp 6399-6408, 2013.
- [6] Kashyap, A.; Joshi, S.D., "Detection of Copy-Move Forgery Using Wavelet Decomposition", International Conference on Signal Processing and Communication (ICSC), India, 12-14 Dec. 2013.
- [7] Hashmi, M.F.; Hambarde, A.R.; Keskar, A.G., "Copy Move Forgery Detection using DWT and SIFT Features", 13th International Conference on Intelligent Systems Design and Applications (ISDA), 2013, Bangi, Malaysia, 8-10 Dec. 2013.
- [8] Hashmi, M.F.; Anand, V.; Keskar, A.G., "A Copy-move Image Forgery Detection Based on Speeded Up Robust Feature Transform and Wavelet Transforms", International Conference on Computer and Communication Technology (ICCCT), Allahabad, India, 26-28 Sept. 2014.
- [9] Xudong Zhao, Shilin Wang, Shenghong Li and Jianhua Li, "Passive Detection of Image Splicing using Conditional Co-occurrence Probability Matrix", 2011 Asia Pacific Signal and Information Processing Association, Xi'an, China, 18-21 Oct. 2011.
- [10] Gajanan K. Birajdar, Vijay H. Mankar, "Blind Authentication of Resampled Images and Rescaling Factor Estimation", 2013 IEEE International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, Pune, India, 15-16 Nov. 2013.
- [11] Archana V Mire, Dr. S. B. Dhok, Dr P. D. Porey , Dr N. J. Mistry, "Digital Forensic of JPEG Images", 2014 IEEE Fifth International Conference on Signals and Image Processing, Jeju Island, Korea, 8-10 Jan. 2014.
- [12] Neenu H.U. , Jini Cheriyan, "Image Forgery Detection based on Illumination Inconsistencies & Intrinsic Resampling Properties", 2014 IEEE International Conference on Magnetics, Machines & Drives (AICERA-2014 iCMMD), Kottayam , India, 24-26 July 2014.
- [13] Shinfeng D. Lin and Tszan Wu, "An Integrated Technique for Splicing and Copy-move Forgery Image Detection", 2011 IEEE 4th International Congress on Image and Signal Processing, Shanghai, China, 15-17 Oct. 2011.
- [14] Muhammad Hussain, Ghulam Muhammad, Sahar Q. Saleh, Anwar M. Mirza, and George Bebis, "Image Forgery Detection Using Multi-Resolution Weber Local Descriptors", 2013 EUROCON, Zagreb, Croatia, 1-4 July 2013.
- [15] Rafael C.G., Richard E.W., "Digital Image Processing", 3rd Edition, Prentice Hall, 2008.
- [16] Tu H.K., Thuong L.T., Synh H.V.U., Khoa H.V., Nga L.T., "Forgery Detection and Identification of the Original and Duplicate Region in Copy-Move Images", The International Symposium on Electrical and Electronics Engineering, HCMC, Vietnam, Oct 30, 2015.
- [17] Kim. S.I., Choi. K.C. , Lee. D.S., "Texture classification using run difference matrix", IEEE 1991 Ultrasonics Symposium, Orlando, FL, Dec 1991.
- [18] Jiyun Yang, Pei Ran, Di Xiao, Jinyong Tan, "Digital Image Forgery Forensics by Undecimated Dyadic Wavelet Transform and Zernike Moments", Journal of Computational Information Systems, No.9, pp 6399-6408, 2013.
- [19] Tu H.K., Thuong L.T., Synh H.V.U., Khoa H.V., "The Efficiency of Applying DWT and Feature Extraction into Copy-Move Images Detection", The IEEE 2015 International Conference on Advanced Technologies for Communications, HCMC, Vietnam, Oct 14-16 2015.



# Portable Facial Recognition Jukebox Using Fisherfaces (Frj)

Richard Mo

Department of Electrical and Computer Engineering  
The University of Michigan - Dearborn  
Dearborn, USA

Adnan Shaout

Department of Electrical and Computer Engineering  
The University of Michigan - Dearborn  
Dearborn, USA

**Abstract**—A portable real-time facial recognition system that is able to play personalized music based on the identified person's preferences was developed. The system is called Portable Facial Recognition Jukebox Using Fisherfaces (FRJ). Raspberry Pi was used as the hardware platform for its relatively low cost and ease of use. This system uses the OpenCV open source library to implement the computer vision Fisherfaces facial recognition algorithms, and uses the Simple DirectMedia Layer (SDL) library for playing the sound files. FRJ is cross-platform and can run on both Windows and Linux operating systems. The source code was written in C++. The accuracy of the recognition program can reach up to 90% under controlled lighting and distance conditions. The user is able to train up to 6 different people (as many as will fit in the GUI). When implemented on a Raspberry Pi, the system is able to go from image capture to facial recognition in an average time of 200ms.

**Keywords**—Facial Recognition; Raspberry Pi; Computer Vision; GNU/Linux Operating System; OpenCV; C++

## I. INTRODUCTION

Facial Recognition is a very difficult topic and computationally strenuous application. However with the advent of newer technology, computers are running faster and faster. New facial detection and recognition techniques are also being developed that are quicker and more reliable. For instance, a revolutionary change to object detection came in 2001 by Viola and Jones when they invented the Haar-based cascade classifier. The accuracy of the Haar can be up to 95% for detecting frontal faces. Still, an even faster technique was the LBP feature detector developed by Ahonen, Hadid and Pietikainen in 2006 [1] [2] [3]. LBP stand for local binary patterns and is potentially several times faster than Haar-based detectors albeit 10-20% less accurate.

Once an object (or face) is detected and an image is preprocessed, there still needs to be a way to train and recognize a new object when presented to the facial recognition system. In 1901 Karl Pearson proposed the technique of Principle Component Analysis, which transforms a set of possibly correlated variables (raw pixels of trained faces) into a smaller set of uncorrelated variables (eigenvectors and eigenvalues). The theory is that in a higher dimensional dataset, most of the information can be described by a few components. These components are called the principal components and are responsible for the most variance in the data [4]. The Eigenfaces representation of faces uses this PCA method to train and store the model used for recognition. However the PCA method has a flaw in that it does not

consider any of the classes (different people) and organizes the principle components purely based on the highest variance the component generates. In the case that an external source (such as light) is generating the variance, the principle components may not contain much discriminative information at all. To combat this issue, another class-specific reduction algorithm was developed by Sir R. A. Fisher and uses Linear Discriminant Analysis (LDA). The method was successfully used to classify flowers in his 1936 paper entitled "The use of multiple measurements in taxonomic problems". In this way, features are found to maximize the ratio of between-classes variation to within-classes variation as opposed to just maximizing overall variation [4]. Thus the recognition algorithm is more robust to external sources such as light. This method is called the Fisherfaces algorithm and is the same algorithm used by the FRJ system.

These advances make it more possible to design real-time portable facial recognition systems, such as the one implemented in this paper. While there are plenty of facial recognition software available in the market today, none provide the feature to play a person's favorite song upon recognition. Also, unlike some systems that use still images to detect faces, our system is a real time system that dynamically detects and recognizes a face depending on the mode of the system [5]. This paper presents a cost-effective solution that is cross-platform (Windows and Linux OS) to train faces of different people and upon recognition of a trained person can either play a song or personalized greeting. The system uses all off-the-shelf components that are easily accessible and does not require special infrared cameras as other systems do [6]. Playing a person's favorite song is useful in boosting a person's morale in much of the same way people use a jukebox to liven a place up. This Facial Recognition Jukebox can also be used to distinguish between different moods (emotion) of a person based on facial expression and play a different song accordingly. Through this emotional recognition it can enable the user or other people in the user's vicinity to become aware of the current user's state of emotion.

Further improvements to facial recognition algorithms still need to be made. With today's technology and algorithms facial recognition is not reliable enough to be used for a true security system. However, they can still be used for other purposes that do not require as high reliability such as playing personalized music or generating a face cartoon [7]. Improvements can be made to make the training set less reliant on lighting conditions or angles in which the person is oriented.

The paper is organized as follows: Section II will present the new FRJ system design, Section III will present the experimental results, Section IV will describe how to use the system, and Section V will present the conclusion remarks.

## II. SYSTEM OVERVIEW

### A. FRJ System Design

The design of the FRJ system can be split into the following major categories: 1. Face Detection 2. Face Preprocessing 3. Face Training 4. Face Recognition 5. Playing Music 6. Saving/Loading Data. A high level block diagram for the code can be seen in Figure 1. The green block illustrates the detection mode. This part of the code is always running by default. First, it acquires a new image and look for a face and eyes. Once the face and eyes are detected there is some preprocessing done to that face region of the image to be used in the next block. The purple Collect Face block code is triggered when a user clicks on the *Add Person* button and will add preprocessed faces for the corresponding person selected. Once the user indicates they are done collecting faces the code continues to the yellow Training Mode block. In this block the Fisherfaces algorithm model is trained using the preprocessed faces and associated face labels (essentially classes). Upon completion of training the code automatically transitions to the red Recognition mode block.

In recognition mode the program basically checks if the captured face matches any of the trained faces and upon matching will play the personalized music of the recognized person. In the event that no person is recognized the code goes back to acquiring the next image and no special action is taken. Side tasks include the ability to save faces so that the user does not need to train a new data set each time. This goes hand in hand with the loading faces functionality being the function to load the previously saved faces. The user can also delete all the faces if they wish to start over in their training. The side tasks are all mouse click based and do not follow the normal process flow as the main functions.

The setup of the overall system can be seen in Figure 2. More detail on the specific components in the setup will be explained in later sections of the paper.

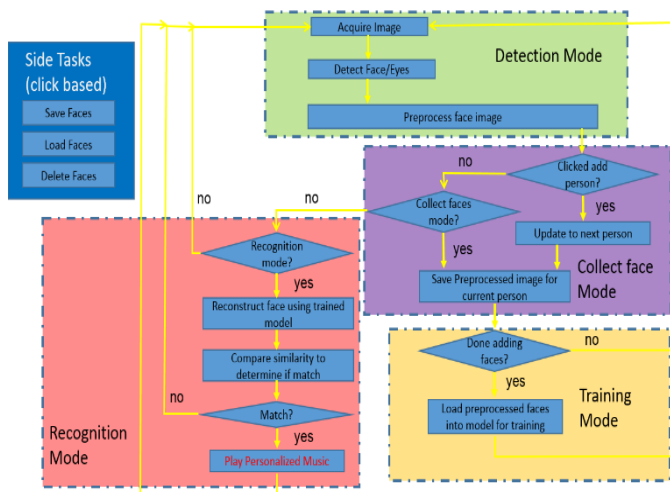


Fig. 1. High Level Block Diagram for Code

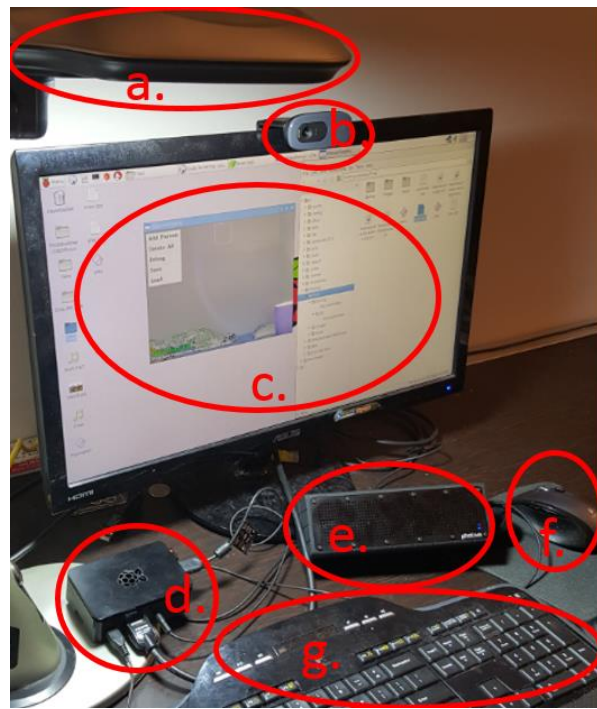


Fig. 2. a. Light fixture b. Webcam c. Display d. Raspberry Pi e. Speaker f. Mouse g. Keyboard

### B. Hardware Used

The computer used in the FRJ system is the Raspberry Pi as seen in Figure 3. It has a 900 MHz quad core ARM Cortex-A7 CPU and a Broadcom VideoCore IV @250 MHz GPU. It contains 1GB of RAM. The reason Raspberry Pi was used is it is a relatively cheap prototyping board and contains enough processing power for the computationally intensive algorithms used for the facial recognition.



Fig. 3. Raspberry Pi Model B: credit card sized mini-computer used to compile and execute C++ code

The Logitech HD Webcam C270 was used as the image capture device as it is UVC compatible, meaning that it is capable of streaming video and has a USB interface. Figure 4 shows the webcam. It captures images at 30 frames/sec. Through software the frame is set to 480 pixels height by 640 pixels width.





Fig. 4. Logitech HD Webcam C270

The rest of the components have some flexibility such as the mouse, keyboard, speaker, and light fixture. The mouse and keyboard are used to interface with the Raspberry Pi Raspbian Linux Operating system to click on buttons as well as type in commands on the command line. The speaker outputs the music file or personalized greeting. Finally, the light fixture is used to control the lighting such that there is strong uniform light on the face during training.

### C. Software Used

Source code was written in C++ as it is a faster computer language for real time systems (as opposed to MATLAB). The OpenCV library was used for the computer vision functions and the Simple DirectMedia Layer (SDL) library was used for playing sound. Both libraries are cross-platform libraries and thus the system is cross-platform compatible over Linux and Windows Operating Systems.

### D. Detection Mode

In detection mode, the program first tries to identify if a face exists within the captured frame. Many pre-trained models are available in the OpenCV official website for the frontal face and eyes. Our program uses the Local Binary Patterns (LBP) classifier for the face (lbpcascade\_frontalface.xml) since it can be a few times quicker than the Haar classifier and thus better for real-time processes. LBP is a little less accurate than the Haar (10-20%) however we will then require detection of both eyes within the face for added reliability.

Once the face is detected, the program will then look for both eyes inside the face to be used in the preprocessing section of the code. Both actual eyes must be detected in order for the preprocessing to work correctly as will be evident in the preprocessing section discussed later. Problems will occur if the eye detector is simply used over the region of the whole face as can be seen in Figure 5. This is because some objects of the face can appear to be eye-like and will be misdetection as eyes, such as the nostril. The solution to the misdetection in eyes is to specify top left and top right regions of the face to search for the left eye and right eye respectively. These regions are based on geometric restrictions in which the majority of human eyes will be located in with respect to the face. Reduction in the search region for the eyes not only reduces the processing time but also increases the reliability of detection of the eyes. Correct detection of the face and eyes can be shown in Figure 6.

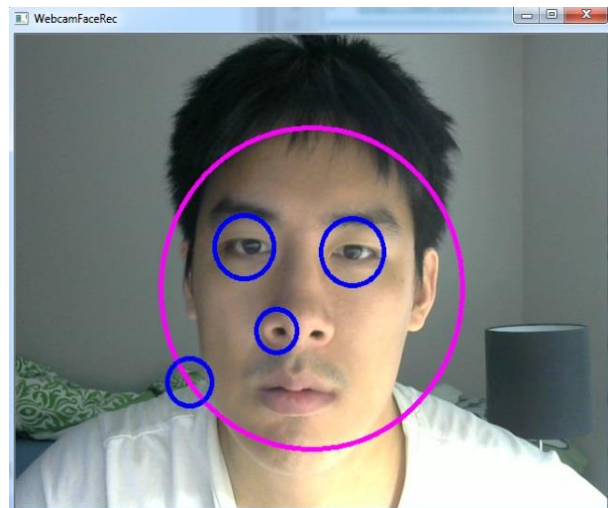


Fig. 5. More than two "Eyes" are detected if the eye classifier is used over the whole face

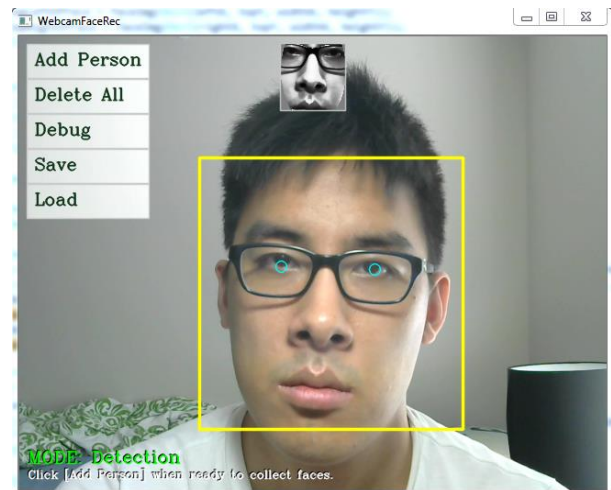


Fig. 6. Correct detection of the face and both eyes

### E. Preprocessing

The preprocessing section is a very important section of the code that conditions the detected face images so that it makes it easier to train the facial recognition model and recognize faces.

First the image is converted to grayscale to reduce the data size as an RGB image would have 3 times the number of pixels. From the detection mode the program is able to detect both eyes within a detected face.

Figure 7 shows the two specified regions in which the program uses to search for both eyes. Once both eyes are detected the program calculates the distance between the eyes and then scales the whole face accordingly such that the distance between the eyes is always the same. The eyes are also adjusted so that they are horizontal and at a specified height.

Finally an elliptical mask is placed on the face to crop out any hair or shadows that may appear on the neck. The end result is a normalized 70 by 70 pixel of a preprocessed face as can be seen in Figure 7.

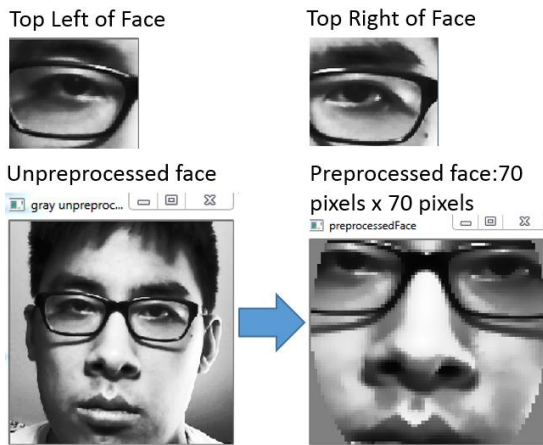


Fig. 7. Preprocessing the face: the position of the eyes are used to scale, rotate, and translate the face

### F. Training

The user is able to initiate training after they are done collecting faces. The training set can be initiated with the adding of new faces per the *Add Person* button or can be loaded with the *Load* button to load faces that were stored previously. An example of a training set is shown in figure 8. The training set should contain different facial expressions, angles and lighting conditions for each person to be robust for recognition later. If the set of faces is trained under strong light on the right side of face and during the recognition mode the user as strong light on the left side of the face, the system will have do a poor job of recognizing the given person. The objective is to get more variation between the different faces so as more information is stored in the principle components. Thus the training set should contain more different conditions to get better results.



Fig. 8. Subset of training set

### G. Recognition

Using the eigenvectors and eigenvalues trained in the model during the training phase, the program then takes the current preprocessed face and projects it into the PCA subspace. It then takes the projection and reconstructs the face into an image again.

If the query image is indeed part of the training set, then the reconstruction should be very good. The program then compares the reconstructed image to the preprocessed image using doing an  $L^2$  relative error norm calculation. A threshold is set for this error. An error calculated that is below the threshold will indicate a match detected. Figure 9 shows an example of a match identified by the program between the face detected and the face of person 2 as can be seen by the green rectangle drawn on person 2's face.

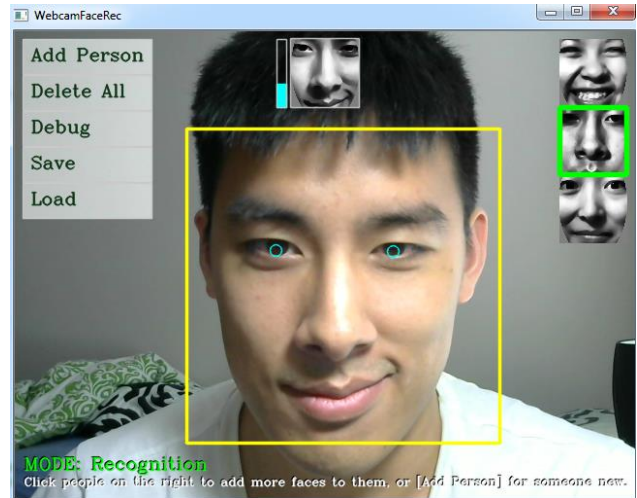


Fig. 9. Example of a recognized face

## III. EXPERIMENTAL RESULTS

Under controlled lighting and distance conditions, the FRJ system is able to achieve up to 90% accuracy in recognizing the correct person. The lamp used was an 1800 lumen lamp with color temperature of 6500 Kelvin. A set distance of  $55 \pm 5$  cm from the webcam was defined. The program can accommodate up to 6 unique people within the GUI for training and recognition.

In terms of timing, two devices were tested for the detection rate and recognition rate, and the speeds can be seen in table 1. The first device was the Raspberry Pi in which our portable program will run. The second device was a benchmarking device and was a Toshiba Laptop computer. Comparing the two, the Toshiba Laptop computer was faster by only 50ms. The Raspberry Pi was still able to go from image capture to recognition within an average time of 200ms. Anything less than 1 second will not be perceivable to the user so using the Raspberry Pi is sufficient.

TABLE I. DETECTION AND RECOGNITION TIMING

Device	Image capture to face detection average time, detection mode (ms)	Image capture to face recognition average time, recognition mode (ms)
Raspberry Pi: Model B	164.05	199.85
Toshiba Laptop: CPU@2.20 GHz, 6GB Insatlld Memory (RAM)	95.5	155.5



#### IV. USING THE FRJ SYSTEM

##### A. Detection mode

No special action is needed for detection mode. The program will automatically start detecting the face and put a yellow rectangle over it and put green circles over the detected eyes. Figure 10 shows an example of detection mode.

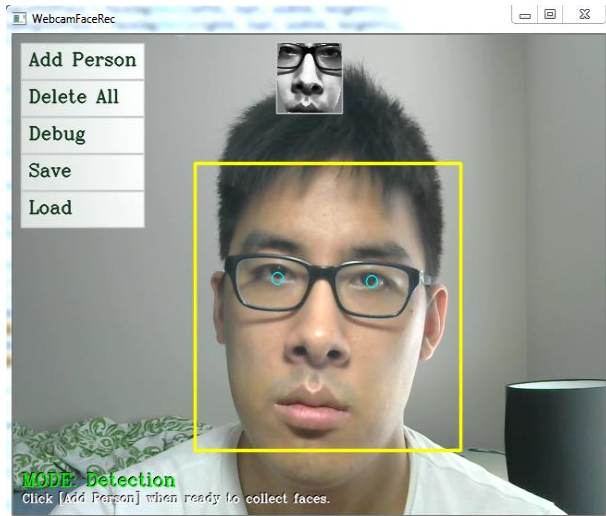


Fig. 10. Example of a recognized face

##### B. Training Faces for the program

To add people to train click on the “Add Person” button. This will add a person to the training set. Each time a picture is taken the frame within the yellow rectangle will flash white to alert the user. Notice that in the Collect Face mode, the person who is getting new faces trained will have a red rectangle over their face. The most recent captured face of each person will be shown on the right edge of the GUI. Figure 11 shows an example of a Collect Face mode face. If the user wants to add more faces to a person already on the screen, the user simply needs to click the most recent picture of said person and then the program will commence collecting faces for that person.

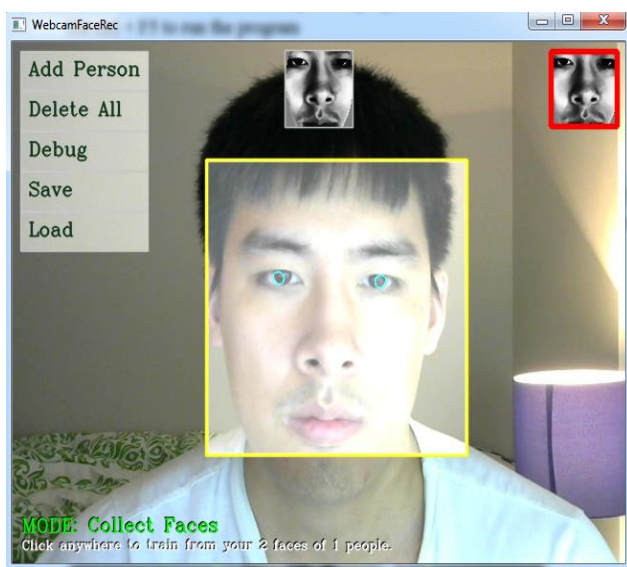


Fig. 11. Example of a face getting collected for the facial recognition model

##### C. Training Mode

After the user has finished collecting faces, the user must click on an area in the GUI not occupied by a face or button. This will commence the training of the preprocessed faces into the model. After the model is finished training the program will automatically switch to recognition mode. Figure 12 shows an example for the training mode.

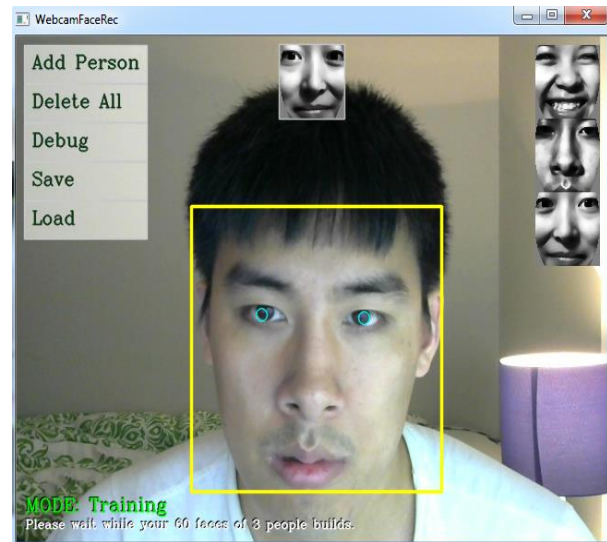


Fig. 12. Faces that were previously collected are now being used to train the model

##### D. Recognition Mode

In recognition mode, once a captured face is recognized with one of the people in the training set a green rectangle will be shown over the recognized person's most recent face and the music file corresponding to that person will play. Figure 13 shows an example for the recognition mode.

The music folder contains the music files as shown in figure 14. The first person corresponds to music file 0.wav. The second person corresponds to music file 1.wav and so on.

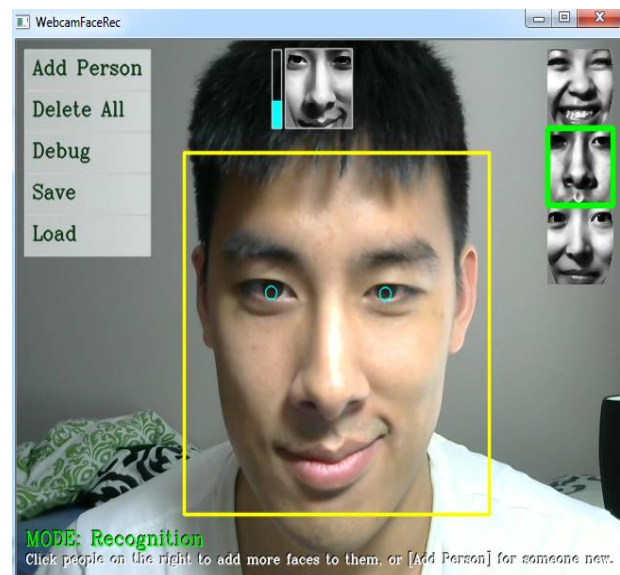


Fig. 13. The correct face is recognized

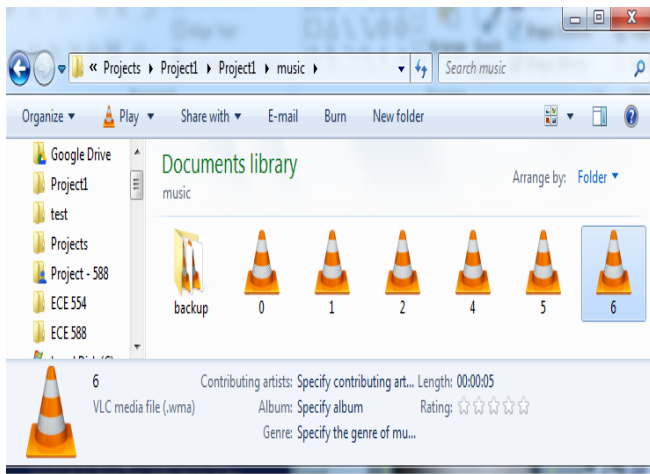


Fig. 14. The music folder

### E. Saving Faces

The program will automatically save faces in the Collect Faces mode into the .bmp format. These images will be stored at <working directory>/images/new\_face/ as shown in figure 15.

### F. Loading Faces

Pressing the “Load” button will load the faces that stored in <working directory>/images/

This feature makes it convenient in that the user does not have to train the model each time with each person at startup as these faces were previously saved.

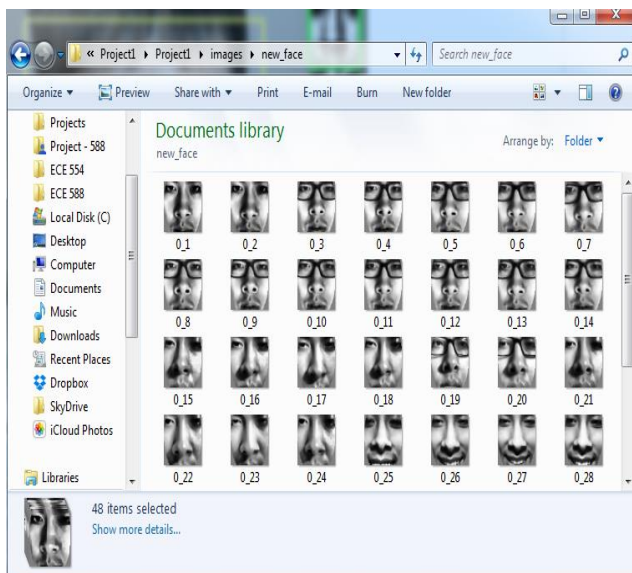


Fig. 15. Collected face directory

### G. Deleting Faces

The user can choose to delete faces and essentially start all over in the training process. To delete all the faces the user must click on the “Delete All” button.

## V. CONCLUSION

A system called the *Portable Facial Recognition Jukebox Using Fisherfaces* (FRJ) was developed and provides a convenient portable machine that plays a person’s favorite song or a personalized greeting upon recognition of a person’s face. The system is implemented on the Raspberry Pi Model B hardware platform and uses OpenCV and Simple DirectMedia Layer libraries for computer vision and media play respectively.

Source code was written in C++. Additional functionality was added to be able to save and load faces trained previously. Currently the system is able to achieve up to 90% accuracy in recognizing the correct trained person under controlled lighting and distance conditions. Up to 6 unique people can be trained as dictated by the size of the GUI. As implemented on the Raspberry Pi, the image capture to facial recognition average time is within 200 ms.

Future work for this project would be to further increase the accuracy of the system by improving upon the Fisherfaces algorithm or by introducing better training mechanisms to be more robust to different environments and people. Also improvements to the HMI/GUI can be made so that it is easier for the user to save and load faces.

## REFERENCES

- [1] D. L. Baggio, “Face Recognition using Eigenfaces or Fisherfaces,” in *Mastering OpenCV with Practical Computer Vision Projects*. Birmingham, UK: Packt Publishing, 2012, ch. 8, pp. 231–268.
- [2] D. Lee et al., “A Face Detection and Recognition System based on Rectangular Feature Orientation,” *International Conference on System Science and Engineering*, 2010.
- [3] Face Description with Local Binary Patterns: Application to Face Recognition, T. Ahonen, A. Hadid and M. Pietikäinen, Proceedings of the IEEE Transactions on PAMI 2006, Vol. 28, Issue 12
- [4] OpenCV Development Team (2015, Feb 25), Face Recognition with OpenCV [Online]. Available: [http://docs.opencv.org/modules/contrib/doc/facerec/facerec\\_tutorial.html](http://docs.opencv.org/modules/contrib/doc/facerec/facerec_tutorial.html), [Accessed October 25 2015]
- [5] P. Laytner et al., “Robust Face Detection from Still Images,” *IEEE Symposium*, 2014.
- [6] M. Weng et al., “Remote Surveillance System for Driver Drowsiness in Real-time Using Low-cost Embedded Platform,” *IEEE International Conference on Vehicular Electronics and Safety*, 2008.
- [7] K. Wang, “Implementation of Face Cartoon Maker System Based on Android,” *Fourth International Conference on Intelligent Control and Information Processing*, 2013.

# Internet of Everything (IoE): Analysing the Individual Concerns Over Privacy Enhancing Technologies (Pets)

Asim Majeed

School of Computing, Telecommunications and Networks  
Birmingham City University (BCU)

Rehan Bhana

School of Computing, Telecommunications and Networks  
Birmingham City University (BCU)

Anwar Ul Haq

Department of Computer Science  
QA Higher Education (ULST)

Imani Kyaruzi

Department of Business Administration  
QA Higher Education (ULST)

Shaheed Pervaz

School of Computing, Telecommunications and Networks  
Birmingham City University (BCU)

Mike-Lloyd Williams

Department of Business Administration  
QA Higher Education (ULST)

**Abstract**—This paper aims to investigate the effectiveness of the provision of privacy of individuals through privacy enhancing technologies (PETs). The successful evolution and emergence of cyberspace with the real world through “Internet of Everything (IoE)” has led to the speedy progress in research and development of predictive analysis of big data. The individual’s privacy has gained a considerable momentum in both industry and academia since privacy-enhancing technologies (PETs) constitute a technical means to protect information. Privacy regulations and state of law deemed this as an integral part in order to protect the individual’s private sphere when the infrastructure of Information Communication Technologies (ICT) is laid out. Modern organisations use consent forms to gather individual’s sensitive personal information for a specific purpose. The law prohibits using the person’s information for purposes other than that of when the consent was initially established. The infrastructure of ICT should be developed in alliance with the privacy laws and made compliant as well intelligent which learn by itself from the environment. This extra layer embedded in the system would educate the ICT structure and help system to authenticate as well as communicate with the perspective users. The existing literature on protecting individuals’ privacy through privacy-enhancing technologies (PETs) is still embryonic and does conclude that the individual’s concerns about privacy are not fully considered in the technological sense. Among other contributions, this research paper will devise a conceptual model to improve individual’s privacy.

**Keywords**—privacy; privacy enhancing technology (PET); big data; information communication technology (ICT)

## I. INTRODUCTION

In recent years, the evolving nature of information systems and the increased processing and storage of personal information in computer databases has made it necessary for ICT practitioners and policymakers to take the issue of

“privacy” more seriously [23],[28]. In particular, the complexity of cloud computing brings a number of known and unknown uncertainties to both service providers and users [12], [19]. The expanding quantity of personal data means that the demand for cloud computing will continue to rise [29]. However, the downside to such developments is the realisation that personal information is constantly recorded and stored without individuals’ consent, therefore, raising a number of concerns. First, the reasons for the collection and the storage of personal information are often neither unknown nor disclosed to the people involved prior to their collection and storage [13]. Secondly, although most software is international, there is no standard mechanism for examining the quality of the databases used to store personal information [9]. Third, there are no uniform ways of handling personal data at the international level and on technical standards, which can help to demonstrate compliance with legal and regulatory frameworks.

Although the term “privacy” seems to have a number of definitions which are sector-specific and tend to carry different meanings depending on varying contexts. The definition that best suits this paper is the one that defines privacy as the right for individuals to be free from secret surveillance and to determine whether, when, how, and to whom, one’s personal or organizational information is to be revealed. According to Guilleaume and Mauree [13] it is suggested that “...Privacy refers to the right to self-determination, that is, the right of individuals to ‘know what is known about them’, be aware of stored information about them, control how that information is communicated and prevent its abuse”.

The meaning of privacy to individuals extends beyond disclosure by suggesting that privacy is also a fundamental human right (see Article 8 of the 1950 European Convention on Human Rights). There are a number of privacy laws and regulations that have been in force since the introduction of the



Internet, however, since then, there has been a number of technological changes – the latest being the privacy challenges brought by the use of “cloud-computing”. In recent years, there have been a number of definitions of cloud computing [23], [28]. In this paper, we adopt Badger et al, [2] definition that describes it as a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources. Sen, [34] simplified it further by suggesting that cloud computing does the network connect a server firm that can host the services for users to it. There is a common consensus in most definitions – all pointing toward the direction that “cloud-computing” exist in a very dynamic environment and hence future definitions are likely to change based on circumstances and contexts [26].

For instance, Pearson, [30] suggests that “the adoption of cloud computing may move quite quickly depending on local requirements, business context and market specificities”. Due to the increased complexities in data collection, storage and dissemination, the economic potential of cloud computing cannot be underplayed [3]. However, there are challenges that seem to accompany the “cloud”. First, there is a notion that the data stored and collected is universal. Pearson, [30] has argued that there are some challenges to providing cloud-computing services including the need to comply with local and regional regulations, obtaining the necessary approvals when data is accessed from another jurisdiction, some additional complexity in terms of governance, maintenance and liability inherent to cloud, and a perceived lack of trust in cloud services. However, along with these challenges, our focus is diverted to the provision of privacy to individuals through privacy enhancing technologies [32].

II. PRIVACY IN THE ERA OF “CLOUD” AND “BIG DATA

Privacy in the era of “cloud” or “big data” has formed an agenda [34]. This is mainly due to the challenges of maintaining trust within a dynamic environment. According to Blaze et al, [4], [38] coping and preserving the privacy of digital identities and challenges are associated with continuous attacks on databases across the world that has forced a number of organisations to deny that their systems have been under attacks [34].

In this paper, our attention is particularly drawn to the “big data” or internally stored data that might be privacy-sensitive – hence the study of Privacy-enhancing technologies (PETs). The privacy enhancing technologies [14] are technologies that protect privacy by protecting personal data and preventing its unnecessary and/or undesired processing but also by making a user aware of the stored data, its processing and the related data flows [13]. PETs are seen as a way to maintain individuals’ privacies by assisting data controllers’ compliance with data protection principles, empowering individuals – by giving them easier access to and control over information about them and allowing them to decide how and when this information will be disclosed to and used by third parties [27].

There are a number of perceived benefits of using PETs. First, this is a cost-effective way of dealing with privacy issues from the onset rather than dealing with privacy and legal compliance issues at a later stage (or once the system is complete). Second, PETs are believed to act as “risk

mitigators” using privacy controls incorporated into electronic information systems to supplement organisational procedures - thus providing additional safeguards to protect individuals’ information from human error [27]. Third, PETs are seen as ways of building by maintaining the integrity of information held [7]. In this paper, four categories of PETs are examined; these include Encryption Tools (e.g., SSL), Policy Tools (e.g., P3P, TRUSTe), filtering Tools (e.g., Cookie Cutters, Spyware) and Anonymity Tools (e.g., Anonymizer, iPrivacy).

A. The Privacy Criterion: Information Life Cycle

The modern data information systems have changed the data protection risks as well as privacy concerns [38]. The new challenges have evolved and ICT can help to minimize and avoid challenges of data protection and privacy [16]. The privacy technologies have been the centre of attention of various researchers since 1970’s. The concerns discussed were refining the privacy principles of identity protection and data minimization through pseudonymisation and anonymization [28]. These discussions led to coin a term “Privacy-Enhancing Technologies (PETs)” considering the full information life cycle from built-in privacy means. The features of data minimisation and privacy by default were stressed and addressed in particular when designing PETs [8].

TABLE I. ADAPTED FROM PISA INFORMATION SECURITY VS PRIVACY [8]

		Privacy Criterion							
		Reporting of processing	Transparent processing	‘As required’ processing	Lawful basis for data processing	Data quality conservation	Rights of the parties involved	Data traffic with countries outside EU	Processing personal data by processor
Information Security	Availability								
	Confidentiality								
	Integrity								

Different organisations implement various rules for measuring the information security and if they are not in compliance to privacy facts of individuals, alternative measures are required to be considered [1]. The attention must be paid to the process of system development especially in relation to those, which may cause privacy hazard in the infrastructure of ICT [22]. The current research is based on the assumption that there is no difference between the methodologies of system development for both private and public environments. Among other researchers, Spiekermann and Cranor, [31] has envisaged an integrated outlook of various techniques and methods with appropriate privacy compliance for the under construction systems. The envisaged model outlines the distinctiveness between Privacy-by-Policy, Privacy-by-Architecture and Privacy-by-Design.

### B. Privacy-by-Architecture

The earliest possible stage of the system development project within information life cycle is intervened by this phase [23]. The technologies are implemented to minimize the personal data collection while keeping it anatomized and protected [20]. The system analyst evaluates the possible dimensions of data breaches at this level and appropriate measures are considered [24]. The procedures and rules are laid out in the form of specifications for the resulting blueprint. The system analyst who would help future users to avoid data breaches inadvertently could avert the pitfalls of programming functionalities.

### C. Privacy-by-Design

The data protection policy makers have proposed the term of “privacy-by-design” [31]. This term was subsequently referred and used in various data protection policies as a recognized recommendation. Privacy-by-design is classified as a sub-part of privacy-by-architecture, which tests the development of Privacy Enhancing Technologies (PETs) at the conceptual paradigm allowing its compliance with ICTs [18]. Various researchers have studied PETs from different security aspects expressing the privacy possibilities for data management aiming to avoid the personal data breach through ICTs. The research led to believe that the users of using those services would be satisfied along with the service providers.

The suggested alternative solution involves interdependent stages since the data precision is decreased after the primary use of information. This solution expresses irreversible way to degrade the data. Among other researchers, Henze et al., [18] that data degradation techniques could be implemented in five possible ways such as, suggests it: upgradable, user-oriented, ability-oriented, service-oriented and external data degradation. All techniques are responsible for built-in system functionality apart from user-oriented data degradation and it is held responsible for the process of data retention [19]. Along with other functionalities these techniques, rely on a single point of interaction to except external data degradation techniques. The entire life cycle of information is self-managed keeping one point of interaction for data degradation and this technique may lead to the privacy solution.

The fundamental issues linked to the implementation of this technique still urge the system architect to predict and ensure all the possible privacy breaches before its execution. These concerns may ensue during the whole lifetime of the ICT system [15]. The information technology is rapidly growing and this assurance is classified as highly contrived. On the other side, the privacy-by-architecture concludes that individuals have no right to say anything about their personal information. The individuals may have their perspective concerns and if they are not heard and managed up to their contents, this approach would be seen as an unwelcome outcome [25].

### D. Privacy-by-Policy

The concept of privacy-by-policy keeps the central rule of “Notice and choice”. The aim of delegating the information processes in the form of notifications, notices and privacy policies are educated to users. The users are flexible to make

choices on their personal data to be used by the organisation on primary or secondary levels. The modern ICT’s infrastructures use this rule as a common practice when deploying policies classified as “choice and consent”. According to Spiekermann and Cranor, [31] this approach is connected with multiple problems of extensive policy documents of privacy and application of incomprehensive applications of millions of users. The privacy approach of “choice and consent” is quite famous within the modern businesses infrastructures, as this does not interfere within the existing layouts of using individual’s personal information extensively [18].

## III. PRIVACY IS CENTRAL TO PETS AND CYBERSPACE

We live in an information society where the use of personal information is constantly forming agendas – mainly, the question is whether since the information is in the open domain is free for everyone to use as they wish [36]. The right of privacy has been well documented in previous studies [27]. There are challenges associated with the use of PETs within the cyberspace environment. First, there is a recurring challenge as to whether computer experts and the technology could be used to protect individuals’ privacy [21]. The answer to this is more complex and demands some empirical evidences.

For example, Rotenberg, [35] has argued that most practitioners tend to use Privacy Enhancing Technologies (PET) that create a technological framework that facilitates the disclosure of personal information, often without any assurance of protection or legal safeguards. He suggests that “these techniques which are often confused with true PETs are put forward by commercial firms and others as a “technical solution” to privacy when in fact they are designed to make it easier to obtain personal data” [29].

### A. ICT and Privacy Enhancing Technologies

The information communication technologies (ICTs) and privacy are often expressed as opponents to each other [33]. The interaction between privacy and ICT is elaborated in this paper exploring the key detail of transmitting the privacy’s conceptual framework to cloud environment. The concerns on individual’s privacy are not new, as they have emerged more during the last half decade of Big Data and Internet of Everything [11]. Various books have been published and researches been conducted on the privacy issues concluding that there is no single rule which complies that the personal information is kept secure. These expressions mean that both new and the existing ICT systems may need to be re-assessed when deploying strategies. The Big Data breakthrough will be adopted into our society during the upcoming years. The continuous capturing of human environment information through sensors embedded within ICT will open new doors of privacy challenges [30].

The personal data of individuals, which used to be stored within organisational ICT systems, would now be residing on clouds in the future. In a traditional way, the term Big Data is associated with the information of users captured and contained by the ICT systems and various analytical tools are used to analyse it, which is the true form its smartness [5]. Information would be kept confidential and private by the

organisations within the era of these technology developments. The privacy infringements implemented within ICT infrastructure are classified as “Law of Nature” which allows them to make choices as well. The organisations establishing ICT infrastructures to process data functionalities may not consider the individuals’ privacy at the development phase but this may be applied in the operational phase.

#### *B. Formation of an ICT system: Privacy-Sensitive Paradigm*

Often with the system based on Internet of Everything there are added constraints in terms of power consumption, limited processing capacity and storage .... (Ibid, 2016) presented a conceptual model (User-driven Privacy Enforcement for Cloud-based Services in the IoE, UPECSI) to address these issues where Internet of Everything interact with cloud based services. One of the key elements of the presented model (UPECSI) was to give user control in a transparent way and the ability to make decisions in privacy settings at varying degrees, instead of accepting a privacy policy at the installation or induction of the service. The presented model was successful in highlighting the need for more control by the user in privacy policy and more user control over degree of services exposed to sensitive private information [10].

However, there seems to be a need for the continuous adaptation of the privacy policy in a dynamic changing environment of Internet of Everything based systems, and simply shifting the focus of privacy policy towards user may not fully address the privacy issues presented by such systems.

In many instances the user themselves would want their privacy policy to intelligently and seamlessly change as the context of use, situation and proximity is changed. In a fast paced dynamically changing and adapting scenario the complete reliance on user driven privacy policy approach may not prove adequate in fulfilling the demands of emerging systems based on IoE infrastructure. To this effect, Artificial Intelligence based approaches could have been injected in the IoE based systems where the Privacy Policy is not only intelligently adaptable but also has a capability to be trained by the user.

Hence, the abstraction layer is suggested for IoE based models where Privacy Policy is presented not as a static component in the system but has adaptable features to inform the services to what degree these can access the private and sensitive information. The training further trigger where

behaviour, context of use, proximity, and situational patterns are transparently allowed to gain access at varying permissible degree and machine assisted technologies then reduce their reliance on user setting this information. This formation would be the case at the start of use of such system when the user would have been more involved in training the intelligent privacy components of IoE based systems.

The ISO certificates provide a measure of compliance for the standards in various sectors such as telecoms, energy, government etc. In 2014, 1,609,294 certificates were issued to management systems across the globe [20]. ISO/IEC 27002, [21] standard provide the framework for establishment of information security management system. The implementation of this framework enable the organisations to systematically preserve integrity and confidentiality of the information and manage the risks related to information security and privacy providing confidence to interested parties on information handling and security of data. The ISO/IEC 27001, [22] is designed to enable organisations to assess, implement and monitor security and privacy issues from internal and external contexts and at different layers of operation and management including understanding of needs and expectations of interested parties at holistic level.

The ISO27002:2013 is based on the guideline of ISO, [22] and framework enable organisations to implement the standard through instruments of control and objectives as provided in table 1. This ISO/IEC 27001: 2013 standard has major influence in directing the security and privacy policies and related structures for major corporates in telecom, service sector and government sector [16].

This framework has been adopted by many organisations around the world (more growth seen in China, India, EU and UK) saw a 7% growth rate from 2013 to 2014 with 23,972 certificates of standard issued by 2014. As the industry is experiencing a new shift towards IoE based systems, the importance of compliance and adherence to standards has become even more important even for small to medium size enterprise. The security and compliance standards needed to evolve in the wake of this shift and incorporate guidelines, measures and controls to keep the trust in the compliance of the standards by the certified organisations. As an example ISO/IEC 27002:2013 standard is discussed in relation to controls which may be needed to add to their existing set of controls for security and privacy compliance.



TABLE II. ADAPTED FROM (GUTIÉRREZ-MARTÍNEZ, [17] AND ISO, [22])

Category	Control	Objective
<b>Policies and regulations of the organization</b>	Information security policies (policies for information security, Review of the policies for information security)	“Direction accordance with business requirements, laws, and regulations.”
	Organisation of information security ((Internal organization: Information security roles and responsibilities, segregation of duties, Contact with authorities, Contact with special interest groups, Information security in project management), (Mobile devices and telework: mobile device policy, Teleworking))	“To control the implementation and operation of information security.”
	Human resources security ((Prior to employment: Screening, Terms and conditions of employment), (During employment: Management responsibilities, Information security awareness education and training, Disciplinary process), (Termination and change of employment: Termination or change of employment responsibilities))	“To protect the organization’s interests ensuring that employees are aware of their information security responsibilities.”
<b>Privacy &amp; Compliance</b>	Asset management ((Responsibility for assets: inventory of assets, Ownership of assets, Acceptable use of assets, Return of assets), (Information classification: Classification of information, Labelling of information, Handling assets), (Media Handling: Management of removable media, Disposal of media, Physical media transfer))	“To ensure that information has an appropriate level of protection. To prevent unauthorized disclosure, modification, removal, or destruction of information stored on media.”
	Access control ((Business requirements of access control: Access control policy, Access to networks and network services), (User access management: user registration and de-registration, user access provisioning, Management of privileged access rights, Management of secret authentication information of users, Review of users access rights, Removal or adjustment of access rights), (User responsibilities: use of secret authentication information), (System and application access control: Information access restriction, Secure log-on procedures, Password management system, Use of privileged Utility programs, Access control to program source control))	“To ensure authorized user access for safeguarding their authentication information.”
	Compliance ((Compliance with legal and contractual requirements: Identification of applicable legislation and contractual requirements, Intellectual property rights, Protection of record, Privacy and protection of personally identifiable information, Regulation of cryptographic controls), (information security reviews: Independent review of information security, Compliance with security policies and standards, Technical compliance review))	
<b>Integrity</b>	Cryptography (Cryptographic controls: Policy on the use of cryptographic controls, key management)	“To protect the confidentiality, authenticity, and/or integrity of information.”
	Physical and environmental security ((Secure areas: Physical security perimeter, Physical entry controls, Securing offices, rooms and facilities, Protecting against external and environmental threats, Working in secure areas, Delivery and loading areas), (Equipment: Equipment siting and protection, Supporting utilities, Cabling security, Equipment maintenance, Removal of assets, Security of equipment and assets off-premises, Secure disposal or re-use of equipment, Unattended user equipment, Clear desk and clear screen policy))	“To prevent loss, damage, theft, or compromise of assets and interruption to the organization’s operations.”
	Operations security ((Operational procedures and responsibilities: Documented operating procedures, Change management, Capacity management, Separation of development, testing and operational environments), (Protection from malware: Controls against malware), (Backup: Information Backup), (Logging and monitoring: Event logging, Protection of log information, Administrator and operator logs, Clock synchronisation), (Control of operational software: Installation of software on operational systems), (Technical vulnerability management: Management of technical vulnerabilities, Restrictions on software installation), (Information systems audit considerations: Information systems and audit controls))	“To ensure correct and secure operations of information processing facilities and to protect against loss of data.”
	System acquisition, development, and maintenance ((Security requirements of information systems: Information security requirement analysis and specification, Securing application services on public networks, Protecting application services transactions), (Security in development and support processes: Secure development policy, System change control procedures, Technical review of applications after operation platform changes, Restrictions on changes to software packages, Secure development environment, Outsourced development, System security testing, System acceptance testing), (Test data: Protection of test data))	“To ensure that information security is designed and implemented across the entire lifecycle of information systems.”
	Supplier relationships ((Information security in supplier relationships: Information security policy for supplier relationships, Addressing security within supplied agreements, Information and communication technology supply chain), (Supplier service delivery management: Monitoring and review of supplier services, Managing changes to supplier services))	“To ensure protection of information that is accessible by suppliers.”
<b>Authenticity</b>	Information security incident Management (Management of information security incidents and improvements: Responsibilities and procedures, Reporting information security events, Reporting information security weaknesses, Assessment of and decision on information security events, Response to information security incidents, Learning from information security incidents, Collection of evidence)	“To ensure a consistent and effective approach to the management of information security incidents, including communication on security events and weaknesses.”
<b>Policies and regulations of the organization</b>	Information security aspects of business continuity management ((Information security continuity: Planning information security continuity, Implementing information security continuity, Verify, review and evaluate information security continuity), (Redundancies: Availability of information processing facilities))	“Continuity in information security management should be integrated into the master plan of the organization.”

#### IV. ANALYSIS AND DISCUSSION

The business models are transforming on a massive scale and this has changed their environment in which they exist and operate. This change has not left service providers free in their actions. The businesses and organisations are moving onto social media and a lot stress and emphasize has been accosted on the user's personal information exploitation. Millions of users have enrolled on various social media sites disclosing their personal information of highly sensitive nature unwittingly. The predictive analytic tools and cloud environments do not provide the privacy like PET integrate. The service providers could facilitate the users' demands and in return, the organisations could make huge revenue. The models of processing the user's personal data are more dependent on investments in cyberspace and PET. The internet companies have evaluated the sensitive information as an instrumental key to their success.

Although the businesses are operating in a model to provide privacy to the users' information but still the users have no obligations to give up their privacy and in this case, businesses would remain profitable. It is doubtless that the privacy and consent models would pose high threats on the business models as well as performances. This increased exposure of ICT as the way to privacy related problems forgoes the idea that it may well not resolve the entire problem, as cultural and social features are inextricably intertwined from the users' perspective. The privacy enhancing technologies defy the personal factors at a sufficiently elaborated level. This also leads information-processing negotiation from the consent perspective especially within the virtualised and cloud based environments.

The parties involved within the privacy awareness of information management have formed various concerns for individual privacies protection laws. The structural movement of cloud-based environments to service-oriented ICTs from ownership-oriented has made the individual privacy resolution almost impossible. The technological and legal aspects are well established and attended by PET but it deals with only certain parts of the ICT infrastructure. While looking at the existing situation from the real life perspective, it is assumed that the responsibility not only depends on ICT infrastructures but also assignable to user behaviour and their consents. The subject is not fully adhered to even after the implementation of procedures of privacy-by-architecture and privacy-by-design. The problem persists and shared between the development methods of ICT infrastructure and individuals who may be thinking that their information would be accessed by whom.

The social interactions are classified as highly volatile source of exploiting the personal sensitive information, which concludes that privacy is extremely implicit and inconclusive and that an ICT infrastructure as well as PET may not be able to defy the subtleties appropriately. Although ICT has evolved tremendously within the current age but still given the current state of technology, it is not feasible to asset pressures on system analyst and developers to exonerate the systems to express the privacy concerns. The development of PET should be moved onto the new conceptual frameworks of privacy-audited and privacy-aware systems. It is envisaged that instead

of yielding blind trust, it is always better to go with informed consent when individuals establishing connections with ICT infrastructures through PET and disclose personal sensitive information.

##### A. Envisaged Model Meets New Privacy challenges

As Internet-based tracking and profiling technologies increasingly expand the ability for e-commerce vendors to collect, store, process and exploit personal data, privacy concern has been identified as a major factor hindering the growth of technologies to protect those. The concerns centre on the confidentiality of accumulated individual personal information and potential risks that individuals experience over the possible breach of confidentiality. The need to protect privacy has led to many initiatives, some behavioural and some technical. Behavioural initiatives generally include providing assurances through privacy seals, government regulations, or addressing individuals' concerns for information privacy, which have been shown to affect trust. While these approaches to protecting privacy are interesting, this paper focuses on an IT artefact that provides one technical solution to the online privacy issue. This approach is in line with a recent review of the privacy literature that highlights the need for more design research in the information privacy domain.

Information is a valuable source and most modern businesses rely on effective use of information for their processes, market reach, customer satisfaction and competitive advantage [9]. This demand for the valuable information puts strain on privacy and data related to personal liking, disliking, and behaviour. Etc. Information system has brought huge success to businesses in achieving their goals. The information system gathers process, distribute, utilise and interact with information [6]. The success of information systems is dependent on channelling communications effectively between different components of such system including people. The information security is an established discipline and with well-defined procedures and measures to this effect.

The Internet of Everything IoE is infusion and interconnectedness of information systems, ICT services/devices and sensor technologies resulting in vast amount of data constantly being generated and updated constantly [29]. This transformation is beginning to break the norms and new systems are based on IoE and are increasingly becoming part of our daily lives, for example smart watches, health and activity monitors. The proximity based services provided by apps using geolocation sensors, remote controlling of home heating system, and intelligent sensors in vehicles, smart rail tickets– the list is endless in many field stretching from leisure, medicine to transportation. The existing security and privacy practises are ill equipped to meet their objectives in the wake of this new shift from information age to the age of IoE [29]. IoE present endless opportunities for the malicious exploitation of such systems e.g. a connected house on low energy consumption might suggest to a hacker that the property is vacant and this information could be used maliciously [37].

The privacy data by the very nature is valuable in information age society, people are increasingly aware of this, and increasingly aware that without their explicit consent the modern system extracts their personal information and

consumes to improve and target their services intelligently. The trend and benefits Internet of Everything brought and highlighted individual privacy concerns as a major obstacle in successful adoption of Internet of Everything as part of living experience at a wider scale [18]. Shifting the balance of privacy settings to individual user add complexity to the design and add burden on the individual user for the understanding and awareness of choices they make and related implication when opting for particular privacy choices or configurations. In many cases individual users are not fully aware of technical complexity of the system and processes in relation to privacy implications [18]. Finding the right balance between system centric and individual centric is a typical dilemma designers of the system face and this problem is exacerbated with the Internet of Everything thus adding complexity and points of pressure in the system in term of making decision for such issues.

## V. CONCLUSION

The theme of this paper builds to form the basis of a dynamic ICT infrastructure, which helps individuals to be connected with each other while keeping the privacy of their personal information. Various searches been performed on databases to reveal that the privacy-awareness within ICT systems is still embryonic and various individual privacy aspects has not as yet been explored. The concept of individual privacy is expressed through laws and rules for organisations to inject privacy-aware concepts within their infrastructure. The professional and scientific committees have paid much attention on the development of various aspects of ICTs and it seems that the right of personal information privacy is lost within the boundaries of organisational amalgamations of laws and technological awareness.

In our view “Intuitive”, privacy and ICT privacy policies are clearly at odds, but legislators, service providers and the public concur in valuing privacy as essential to acceptance of information technology-based services. Providing proper privacy to individuals is therefore no matter of small concern. Making clear to all parties involved that their respective responsibilities cannot be delegated to ICTs is crucial. The infrastructure of ICT is developed in alliance to the privacy laws and made compliant as well intelligent which learn by itself from the environment. This extra layer embedded in the system would educate the ICT structure and help system to authenticate as well as communicate with the perspective users. Governmental, service providers and individuals’ concerns should be properly addressed to retain the privacy levels that form the essence of civil liberties and maintain freedom in society. To create a truly privacy-aware ICT, a holistic approach is needed in finding methods to shift control over information back towards the individual. Taking the ICT from an individual’s perspective as a starting point would allow for a first step towards a true impact analysis of ICTs on what is considered a building block of free societies.

## REFERENCES

- [1] Anderson, A. (2005). A Comparison of Two Privacy Policy Languages: EPAL and XACML: Sun Microsystems Laboratories. from <http://research.sun.com/techrep/2005/sml-tr-2005-147/TRCompareEPALandXACML.html>.
- [2] Badger, L., Grance, T., Patt-Corner, R., & Voas, J. (2011). Draft Cloud Computing Synopsis and Recommendations. National Institute of Standards and Technology (NIST) Special Publication 800-146. US Department of Commerce. May 2011. Available online at: <http://csrc.nist.gov/publications/drafts/800-146/Draft-NIST-SP800-146.pdf> (Accessed on: February, 2016).
- [3] Birman, K., Chockler, G., & van Renesse, R. (2008). Towards a cloud computing research agenda, [http://www.cs.cornell.edu/projects/quicksilver/public\\_pdfs/SIGACT2.pdf](http://www.cs.cornell.edu/projects/quicksilver/public_pdfs/SIGACT2.pdf).
- [4] Blaze, M., Kannan, S., Lee I., Sokolsky, O., Smith, J. M., Keromytis, A.D., & Lee, W. (2009). Dynamic Trust Management. IEEE Computer, Vol 42, No 2, pp. 44-52, 2009.
- [5] Bennett, J. and S. Lanning (2007). The Netflix prize. In Proceedings of KDD Cup and Workshop.
- [6] Beynon-Davies, P., 2013. Business Information Systems. 2nd edition ed. s.l.:Palgrave Macmillan.
- [7] Bowers, K., Juels, A., (2009), Oprea, A.: HAIL: A high-availability and integrity layer for cloud storage. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, CCS’09, pp. 187–198. ACM, New York, NY, USA (2009). DOI 10.1145/1653662.1653686
- [8] Borking, J. (2010) Privacyrecht is code. Over het gebruik van Privacy Enhancing Technologies, Deventer: Kluwer.
- [9] Buyya, R., Yeo, C., & Venugopal, S. (2008). Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities. Retrieved January 21, 2016 from author’s website: [http://www.buyya.com/papers/hpcc2008\\_keynote\\_cloudcomputing.pdf](http://www.buyya.com/papers/hpcc2008_keynote_cloudcomputing.pdf)
- [10] Cannon, J. C. (2005). Privacy: What Developers and IT Professionals Should Know: Addison-Wesley.
- [11] G. Aggarwal, E. Bursztein, C. Jackson, D. Boneh, (2010), An Analysis of Private Browsing Modes in Modern Browsers, In: Proceedings of 19th USENIX Security Symposium, Washington, DC, USA, pp. 79–94.
- [12] Garfinkel T, Rosenblum M., (2005), When virtual is harder than real: Security challenges in virtual machine based computing environments. In Proceedings of the 10th conference on Hot Topics in Operating Systems, Santa Fe, NM. volume 10. CA, USA: USENIX Association Berkeley; 2005:227–229.
- [13] Guilloteau, S and Mauree, V (2012) Privacy in Cloud Computing ITU-T Technology Watch Report March 2012 Retrieved from: [https://www.itu.int/dms\\_pub/itu-t/oth/23/01/T23010000160001PDFE.pdf](https://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000160001PDFE.pdf)
- [14] G.W. Van Blarckom, J. B. (2003). Handbook of Privacy and Privacy-Enhancing Technologies - The case of Intelligent Software Agents. Retrieved from e-Europe: <ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/DPP/CWA15263-00-2005-Apr.pdf>
- [15] Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In Proceedings of the 41st annual ACM symposium on Theory of Computing, pp. 169–178.
- [16] Gartner, (2013). ISO/IEC 27001:2013 Shifts Focus From the Effectiveness of Controls to Risk Treatment Plans. [Online] Available at: <https://www.gartner.com/doc/2630435/isoiec--shifts-focus-effectiveness> [Accessed 14 February 2016].
- [17] Gutiérrez-Martínez, J., Núñez-Gaona, M. A. & Aguirre-Meneses, H., (2015). Business Model for the Security of a Large-Scale PACS, Compliance with ISO/27002: 2013 Standard. Journal of digital imaging, 28(4), pp. 481-491.
- [18] Henze, M. et al., (2016). A comprehensive approach to privacy in the cloud-based Internet of Everything. Future Generation Computer Systems, pp. 701 -718.
- [19] Hashizume K, Yoshioka N, Fernandez EB., (2013), Three misuse patterns for Cloud Computing. In Security engineering for Cloud Computing: approaches and Tools. Edited by: Rosado DG, Mellado D, Fernandez-Medina E, Piattini M. Pennsylvania, United States: IGI Global; 2013:36–53.

- [20] ISO, (2014). iso\_survey\_executive-summary. [Online] Available at: [http://www.iso.org/iso/iso\\_survey\\_executive-summary.pdf?v2014](http://www.iso.org/iso/iso_survey_executive-summary.pdf?v2014) [Accessed 14 February 2016].
- [21] ISO, (2016). ISO/IEC 27001:2013(en) Information technology — Security techniques — Information security management systems — Requirements. [Online] Available at: <https://www.iso.org/obp/ui/#iso:std:iso-iec:27001:ed-2:v1:en> [Accessed 14 February 2016].
- [22] ISO, (2016). ISO/IEC 27002:2013 (en) Information technology — Security techniques — Code of practice for information security controls. [Online] Available at: <https://www.iso.org/obp/ui/#iso:std:iso-iec:27002:ed-2:v1:en> [Accessed 14 February 2016].
- [23] Jaeger, P.T., Lin, J., and Grimes, J. (2008). Cloud Computing and Information Policy: Computing in a Policy Cloud? *Journal of Information and Politics*, 5(3): 269-283.
- [24] Jensen, M., Gruschka, N., Herkenhoner, R., (2009), A survey of attacks on web services. *Computer Science – Research and Development* 24(4), 185–197.
- [25] Kobsa, A., & Teltzrow, M. (2005). Contextualized Communication of Privacy Practices and Personalization Benefits: Impacts on Users' Data Sharing Behavior.
- [26] D. Martin & A. Serjantov (2004.), *Privacy Enhancing Technologies: Fourth International Workshop, PET 2004*, Toronto, Canada (Vol. LNCS 3424, pp. 329-343). Heidelberg, Germany: Springer Verlag.
- [27] Madhub, D (2012) *PRIVACY ENHANCING TECHNOLOGIES An Absolute Necessity for Effective Compliance with Data Protection Laws*. Available at: [http://dataprotection.govmu.org/English/Documents/Publications/Guidelines/DPO\\_Vol7\\_EnhancingTechnology.pdf](http://dataprotection.govmu.org/English/Documents/Publications/Guidelines/DPO_Vol7_EnhancingTechnology.pdf) (February, 2016)
- [28] Mell, P., & Grance, T. (2009). NIST Definition of Cloud Computing. Retrieved from NIST [www.nist.gov/itl/cloud/upload/cloud-def-v15.pdf](http://www.nist.gov/itl/cloud/upload/cloud-def-v15.pdf)
- [29] Maras, M.-H., (2015). Internet of Everything: security and privacy implications. *International Data Privacy Law*, 5(2), pp. 99 - 103.
- [30] Pearson, S., Benameur, A., (2010), Privacy, security and trust issues arising from cloud computing. In: *Proceedings of the International Workshop on Cloud Privacy, Security, Risk, and Trust, CLOUDCOM'10*, pp. 693–702. IEEE Computer Society, Washington, DC, USA (2010). DOI 10.1109/CloudCom.2010.66
- [31] Spiekermann, S. and Cranor, L.F. (2009) 'Engineering privacy', *IEEE Transactions on software engineering*, vol. 35, no. 1, pp. 67-82
- [32] PrivacyChoice, PrivacyChoice Tracking Protection List, 2010.
- [33] Richards, K. and E. Jones (2008). Customer relationship management: Finding value drivers. *Industrial Marketing Management* 37(2), 120–130.
- [34] Sen, J. (2011a). A Robust Mechanism for Defending Distributed Denial of Service Attacks on Web Servers. *International Journal of Network Security and its Applications*, Vol 3, No 2, pp. 162-179, March 2011.
- [35] Solove, D.J., Rotenberg, M., and Schwartz, P.M. (2006) *Privacy, information, and technology*, New York: Aspen Publishers Online.
- [36] Tsai, J., Egelman, S., Cranor, L., & Acquisti, A. (2007). The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study. In *proceedings of the Sixth Workshop on the Economics of Information Security*, Pittsburgh, PA.
- [37] Want, R., Schilit, B. N. & Jenson, S., (2015). Enabling the Internet of Everything. *Computer*, Volume 1, pp. 28 - 35
- [38] Zetter, K. (2010). Google hackers Targeted Source Code of More Than 30 Companies. *Wired Threat Level*. January 13 2010. Available online at: <http://www.wired.com/threatlevel/2010/01/google-hackattack/> (Accessed on: February, 2016).

# A General Evaluation Framework for Text Based Conversational Agent

Mohammad Hijjawi

Faculty of Information Technology  
Applied Science Private University  
Amman, Jordan

Zuhair Bandar

School of Computing  
Manchester Metropolitan University  
Manchester, UK

Keeley Crockett

School of Computing  
Manchester Metropolitan University  
Manchester, UK

**Abstract**—This paper details the development of a new evaluation framework for a text based Conversational Agent (CA). A CA is an intelligent system that handle spoken or/and text based conversations between machine and human. Generally, the lack of evaluation frameworks for CAs effects its development. The idea behind any system’s evaluation is to make sure about the system’s functionalities and to continue development on it. A specific CA has been chosen to test the proposed framework on it; namely ArabChat. The ArabChat is a rule based CA and uses pattern matching technique to handle user’s Arabic text based conversations. The proposed and developed evaluation framework in this paper is natural language independent. The proposed framework is based on the exchange of specific information between ArabChat and user called “Information Requirements”. This information are tagged for each rule in the applied domain and should be exist in a user’s utterance (conversation). A real experiment has been done in Applied Science University in Jordan as an information point advisor for their native Arabic students to evaluate the ArabChat and then evaluating the proposed evaluation framework.

**Keywords**—Artificial intelligence; Conversational Agent and evaluation

## I. INTRODUCTION

Different terms can be used to define a system has the ability to handle user conversations such as Conversational Agent (CA), dialog system and chatterbot. CAs are playing significant roles in different applications, for instance, in marketing, education, help desk, entertainment, e-commerce, information retrieval and generally in business [1].

Basically, it can be considered that the first try to build a CA was in 1950 by Alan Turing and it called the imitation game or Turing test [2]. Turing test tried to determine if a computer program could think or at least imitating a human behaviour. In the Turing test, an interrogator sends a series of teletype questions to a hidden participant through a computer link. Then the interrogator has to distinguish whether the hidden participant is human or machine based upon the hidden participant’s replies [2].

Comparing what Turing expected in his article [2] and what we have today, we could conclude that Turing’s expectations have not been met. Although, it is now more than sixty years since Turing stated his beliefs and despite the fact that computer storage capacities exceeded his request (1 GB), no computer program has been able to pass the Turing test(imitation game) successfully [3, 4]. According to [5], in

order for a computer program to pass Turing test, the machine must acquire the same level of intelligence as a human in all cognitive tasks. However, since the first CA (imitation game or test [2]) which Alan Turing tried to make the machine to have chatting with human, several types has been raised. These types targeted different kinds of conversations starting from text typing conversation, spoken conversation and mixed among them conversation. Due to this difference, number of approaches has been proposed and used to develop CAs which are Natural Language Processing (NLP), Pattern Matching (PM) and Semantic Sentence Similarity (SSS) measures.

All of the CA’s building approaches (except PM) are natural language dependant which means it needs to understand the targeted natural language. Understanding and computing the natural language is quit complex and needs different huge research effort from the language scholars before. Given this, most of CAs has been developed using the PM approach for its simplicity and as it natural language independent. In addition, the pattern matching is not expensive computationally as it does not need a complex pre-processing steps to understand the sentence (user’s utterance). Consequently, number of CAs such as ArabChat [1], InfoChat [6], ELIZA [7] and ALICE [8] used this approach to handle conversations for applications deal with large numbers of users in a real-time environment like the Internet [1]. Basically, this approach (the PM) based on matching a conversation with a pre-structured patterns to find the suitable one. Then, the response that related to the best matched pattern will be replied [1]. The NLP which is defined in computing as “the computational processing of textual materials in natural human languages” [9] is based on understanding a sentence. Technically, NLP based CAs uses grammar rules and a list of attribute/value pairs to extract the conversation’s speech act type from the sentence [10]. Then, it use these extracted information to fill a template-based response [10]. However, extraction such information is not easy at all as it depends on many linguistic factors [10]. In a rich language especially the sematic languages such as Arabic, this extraction will be harder to process [3, 10]. The SSS approach is based on checking the similarity level in semantic between two sentences [11]; the first sentence is the conversation itself and the second is a scripted pattern inside the CA. the most closed pattern in semantic(meaning), its response will be replied as an answer to the conversation. The SSS approach is based on computational semantic based manual built databases such as WordNet [11, 12]. However, such database has been established in 2006 [11] and the

research in SSS in general is still a young research area in the Arabic language [3, 12].

All of these approaches (NLP, PM and SSS) has different advantages and disadvantages as discussed in [3, 10-12]. However, most of the evaluation methodologies for these approaches has been done manually by checking the CA's logs [3]. Evaluating CAs is not an easy task as it depends on number of factors are not easy to measure [3, 4].

Generally, there are many types of systems that deal with text based sentences formed in a specific natural language, such as Information Retrieval (IR), Natural Language Processing (NLP), Question Answering (QA) and Conversational Agents systems. However, the evaluation process of these systems varies due to the differences in their working mechanism and their output. When researchers of IR systems want to evaluate their system, they might be interested in ranking the returned documents according to the entered keywords. Therefore, they usually use special metrics for evaluation, such as the "recall" (the percentage of retrieved documents that are relevant) and "precision" (the percentage of relevant documents that are retrieved) metrics [13]. NLP systems might be evaluated by comparing the output with a prebuilt perfect result document, called the "gold standard". Usually, QA systems use the same metrics as IR systems ("recall" and "precision" metrics). QA systems are IR systems with an extra processing module to analyse the retrieved documents and extract a response [13].

Applying such evaluation techniques to evaluate CAs may not be useful due to the differences in their working mechanism and their output. Although QAs and CAs have the same output (response for the entered utterance), they differ in their working structures.

There exists two primary approaches to evaluating CAs: objective and subjective evaluation approaches. The objective approach can be done without recourse to human judgment. This approach is based on systematic and scientific measures to evaluate a CA [14]. Conversely, the subjective approach can be performed with a recourse to human judgment by asking him/her about his/her opinion of using a CA [15].

The subjective approach usually utilises a user questionnaire to evaluate the CA. This questionnaire might be used to ask the user (after using the CA) about several aspects of the CA such as the CA usability, naturalness or his/her overall satisfaction of using the CA. However, it is impossible to rely on user to give his/her opinion regarding CA internal components' performance as he/she has no idea about them. A special type of a CA evaluation, based on human (judges) to determine the most human-like CA among competitors, is the Loebner Prize competition [16]. In 1990, the Loebner Prize was established in collaboration with the Cambridge Centre for Behavioural Studies. This prize aims to encourage researchers to develop Conversational Agents. The Loebner competition uses expert human judges to evaluate the competing CAs using the Turing test. Passing the Turing test means that the program's responses should be indistinguishable from human conversation. This method of evaluation is not academically rigorous, and not all CAs can participate. Since the Loebner competition was established,

many Conversational Agents have competed for it and not one CA has passed the Turing test. Unfortunately, some of these Conversational Agents focus merely on passing the test, rather than on advancing the field of Conversational Agents [4].

The objective evaluation approach evaluates a CA as a whole system (black box approach) or evaluates the CA's components individually (glass box approach). A black box evaluates the system as a whole, based on user satisfaction. This is usually done by evaluating inputs and outputs without considering any internal details [17]. The black box focuses on the performance of the system in terms of number of aspects, such as achievement task and the cost of that achievement [17]. The glass box deals with internal details by evaluating the individual components of a system [18]. An example of the glass box approach to evaluation is undertaken to measure the error rate on the sentence recognition module which is included in a spoken CA called ARPA [18]. A black box evaluation approach was used to evaluate the SUNDIAL CA [19]. This approach determined the SUNDIAL's user satisfaction by determining the task and its cost. The cost of the determined task might be based upon number of utterances needed to achieve the task, the elapsed time to complete the task and the quality of interaction among conversation entities.

PARADISE [14] (PARADigm for DIAlogue System Evaluation), is a framework used for evaluating spoken CA. PARADISE relies on a comparison between agents through achieving the maximum user satisfaction.

Maximum user satisfaction means maximum task success with the minimum cost. PARADISE measures the task success per dialogue or sub-dialogue by determining the information requirement needed to exchange between the agent and the user. This information, compared with a prebuilt confusion matrix, is collected via controlled experiments for these agents that "summarizes how well an agent achieves the information requirements of a particular task for a set of dialogues instantiating a set of scenarios" [14].

PARADISE calculates the task cost by measuring two factors: firstly, task efficiency, which might be represented through determining number of utterances that takes to complete the task and the elapsed time that it needs; secondly, measuring the quality of the task, which might be determined, based upon the agent response delay and utterances' recognition errors rate (spoken utterances). PARADISE considers a small number of the total utterances needed to achieve the task better than a large number. This might be true with a CA that provides information for a train schedule between cities, for example. In contrast, this might be not true for other CAs that are designed to handle open conversations (the user converses in general about the selected domain's topic) between the CA and user. Therefore, a CA that considers the largest number of the total utterances might perform better, assuming that a larger number might mean that the user is more interested in using the CA.

Evaluating a CA is a divergent problem due to the number of metrics that can be used to evaluate it. For instance, a CA can be evaluated using usability metric [20, 21], user satisfaction metric [14, 22], response quality metric [23], ease of use metric [15], conversation duration metric [24], task

completion level metric [25] and natural agent behaviour metric [15]. Each of these metrics has its own characteristics, objectives and its techniques and thus using all of these metrics might be not useful for evaluating a specific CA. According to [26], the best CA evaluation should be related to the nature of the CA's task and the users' needs. For instance, evaluating a ticket booking CA differs from evaluating a psychiatrist CA. The fundamental purpose behind a system's evaluation is to improve its performance. The lack of a comprehensive evaluation framework has been a limiting factor in the growth of Conversational Agents [4]. In addition, different CAs might need different approaches to evaluate [26].

a CA evaluation plays an important role for all participants building and using the CA [27]. It is important for CA developers "to tell if their system is improving", and for CA's integrators "to determine which approaches should be used where", and for consumers also "to identify which system will best meet a specific set of needs" [27]. Therefore, a combination of objective measures and subjective measures will be better for evaluating a Conversation Agent.

## II. THE SELECTED CASE STUDY ARABCHAT

In this paper, a specific text based CA called ArabChat [3] has been chosen to evaluate the proposed evaluation methodology. The ArabChat is a related research work for the paper's authors so it is easy to access and this is the reason why it has been chosen in this research. The ArabChat is an Arabic based CA which means it handles Arabic conversations. This is the reason why the proposed and developed evaluation framework in this paper called the "ArabChat Evaluator". However, the developed evaluation methodology can work for any CA for all natural languages which means it is language independent.

The ArabChat uses the Pattern Matching technique to handle the Arabic textual conversations. The development of ArabChat needs to meet three requirements: scripting language, engine and brain. The scripting language will be used to script the specific domain aspects in order to represent them. While, the brain is a structured store or knowledge base that is used to store the domain's scripts. The engine handles user's utterances (conversations) that target the scripted domain.

The ArabChat is a rule-based Conversational Agent and it fundamentally is comprised of a novel scripting engine and a rule-based scripting language structured in a novel way to handle the topics (contexts) of conversations. Each context (main topic) has several rules (sub-topic) and each rule has several patterns (to be matched with user sentence) and responses. Each context has a default rule to be fired when no rules matched a user's utterance.

ArabChat is a turn-based Conversational Agent, which means each one of the conversation's parties (user and ArabChat) has its turn for conversation. Once the user enters his/her utterance, ArabChat processes this utterance and replies with a suitable response. The conversation remains ongoing until one of the conversation's parties terminates it.

The ArabChat was deployed and published in ASU (Applied Science University) in Jordan to work as an information point advisor for their users (registered students, unregistered students and employees).

A comprehensive evaluation methodology consisting of objective and subjective approaches has been used to evaluate the ArabChat. The subjective approach has been done through asking the ArabChat's users about their opinions from different aspects by filling an online questionnaire [3]. Where the objective approach has been conducted through automatic evaluation techniques and manual analysing and consists of the "Glass box" and "Black box" approaches [3]. The "Glass box" approach evaluated ArabChat components individually. The ArabChat obtained a 67.836% of general user satisfaction [3]. This result can give a general overview of ArabChat performance, but it does not give a full indicator about its performance. Hence, the "Black box" approach using the proposed and developed methodology in this paper will be used to evaluate ArabChat and giving more accurate indication.

The ArabChat was evaluated depending on the ratio of matched and unmatched utterances [1]. This technique might give a general overview about the ArabChat's performance. However, it cannot give an accurate result as the utterances might match wrong rules. Therefore, in this paper a new framework will be modelled and developed to evaluate the ArabChat in a more accurate way. The next section is describing the proposed and developed framework.

## III. THE "ARABCHAT EVALUATOR"

The "ArabChat Evaluator" is based on the black box evaluation approach which means testing and evaluating the ArabChat CA as one unit. The "ArabChat Evaluator" is based on a comparison process between the user's utterance and ArabChat's response in terms of the existing of "Information Requirements" (discussed later) words without dealing with any internal component details.

Generally, in a conversation between a user and a CA words need to be exchanged between them. Regardless of the CA type (spoken or textual), these words are found in the user's utterance. In the ArabChat case, these words are in the text form. Not all of the utterance's words are important to check but some of them are important (keywords) to check. For instance, the utterance "من فضلك، ما هو ايميل رئيس الجامعة؟" "Excuse me, what is the university president's email?" contains 7 words. The only important words (keywords) are 3 words which are "ايميل" "email", "رئيس" "president" and "الجامعة" "university" which they construct the topic "Email of university's president". These important words will be called in this evaluation "Information Requirements".

The "ArabChat Evaluator" aims to evaluate the ArabChat's performance through the analysis of the quality of ArabChat's response, which might indicate the user's satisfaction. The quality of a response means how much a replied response is related to the processed utterance.

The "ArabChat Evaluator" is a separate system from ArabChat and works in an offline mode whenever it needs to

evaluate ArabChat. Before proceeding with discussing the “ArabChat Evaluator” methodology, it is important to discuss the “Information Requirements” that “ArabChat Evaluator” is based on.

A. The “Information Requirements”

Each rule in ArabChat has its own topic to handle. An utterance that causes a rule to fire (the utterance matched one of the rule’s patterns) should contain some keywords related to the rule’s topic. For instance, a rule X is designed to reply to users asking about fees of a computing course in ASU. In order to fire the rule X, an utterance should contain at least two keywords, which are “سعر” “fee” and “حاسوب” “computing”, structured in a suitable way in the utterance. These two keywords (“سعر” and “حاسوب”) are considered as “Information Requirements” to fire rule X.

The “Information Requirements” is part of a rule structure and it contain a list of numbers with each number representing a list of different keywords. These keywords are stemmed and grouped semantically in separate groups, as presented in Table 1. This table represents a sample of the whole ArabChat “Information Requirements” list. The first group in Table 1 has 4 words, all of which might convey the same meaning (“fee”). Although, the fourth group has 3 words with 3 different meanings, all of them might be semantically related, and thus they are put in the same group. For the above mentioned rule (X), its “Information Requirements” parameter (according to Table 1) is (1,3) which represent the keywords (“سعر” “fee” and “حاسوب” “computing”). Implicitly, it is possible to consider “Information Requirements” as a list of keywords. ArabChat has list of “Information Requirements” that contain all ArabChat applied domain’s keywords. These keywords or these parameters does not involved in the user’s conversation handling process. As mentioned before, the ArabChat engine is based on the pattern matching technique to handle the conversations.

TABLE I. A SAMPLE OF ARABCHAT “INFORMATION REQUIREMENTS” LIST

#	Group words	Group words in English
1	سعر, ثمن, كلف, رسم	Price, cost, fee
2	مدد, جند, اجل	Delay, postpone
3	حسب, برمج, كمبيوتر	Compute, program, computer
4	تاريخ, وقت, ساعة	Date, time, hour

B. The “ArabChat Evaluator” framework

The “ArabChat Evaluator” is based on the “Information Requirements” that are exchanged between a user and ArabChat in order to evaluate the ArabChat. The “ArabChat Evaluator” works in offline mode and in isolating from the ArabChat. This means that it only used when it needs to evaluate the ArabChat which means it does not affect the ArabChat’s performance when handling users’ conversations.

The mechanism used to determine the “Information

Requirements” of an utterance differs from that used to determine the “Information Requirements” of a response. The mechanism used to determine the “Information Requirements” of an utterance is based on checking the utterance’s words, as it will be described later in this section, while the mechanism used to determine the “Information Requirements” of a response is based on the “Information Requirements” parameter of a rule that belongs to this response. Each response in ArabChat’s domain should belong to a specific rule. Each rule has an “Information Requirements” parameter. The “Information Requirements” parameter is not involved in the pattern matching process that the scripting engine adopted to match an utterance and then fire a rule. As discussed before, this parameter is just used for evaluation purposes, which means it used after all users finish their conversations with ArabChat. In contrast, during a user conversation, ArabChat accumulates the “Information Requirements” parameter contents related to the fired rules and stores them in ArabChat logs. Figure 1 shows the “ArabChat Evaluator” framework.

The “ArabChat Evaluator” reads the contents of the “Brief Log” (located in ArabChat brain) record by record in order to acquire its input (utterance and response) and produce an output (evaluation results). The “Brief Log” has 3 blank fields: “response evaluation”, “Patterns scripting evaluation”, and “conversation evaluation”. These blank fields will be filled by the “ArabChat Evaluator” for each utterance. Filling these fields means evaluating the processed record (utterance). Each record represents a conversation between a user and ArabChat. The “ArabChat Evaluator” starts its work by reading the first unevaluated record (its three fields are blank) in the “Brief Log”. Then, it moves to the next unevaluated record and so on. Table 2 shows a customised sample of the components of the “Brief Log” before the evaluation process begins.

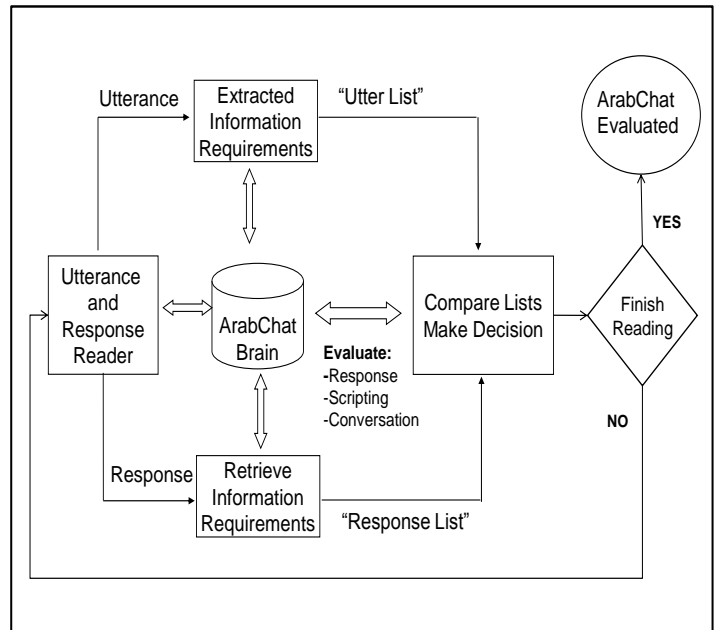


Fig. 1. The “ArabChat Evaluator” framework



TABLE II. A CUSTOMISED SAMPLE OF “BRIEF LOG” COMPONENTS BEFORE THE EVALUATION PROCESS

#	Utterance in Arabic	Utterance in English	Response in Arabic	Response in English	Information Requirement	Response evaluation	Patterns Scripting evaluation	Conversation evaluation
1	كم سعر ساعه تخصص الحاسوب	What is the price of a credit hour of computing course.	سعر الساعة المعتمدة لتخصص علم الحاسوب هو 75 دينار أردني	The price of a credit hour for computing course is 75 JD which is equaled to 106 USD	8,8,48			
2	كم عدد الساعات لتخصص التمريض والمحاسبة	What is the price of a credit hour of nursing accounting.	عدد الساعات المعتمده لتخصص علم الحاسوب هو 132 ساعة معتمدة	Total number of credit hours of Computer Science is a 132 hours.	29,29			
3	لقد سمعت من أصدقائي بأنك ثرثار	I have heard from my friends that you are talkative	تقدم جامعة العلوم التطبيقية العديد من التخصصات في شتى المجالات وقد صمم هذا النظام ليقيم بالتسهيل عليك عملية الاستفسار والردشة مع موظفي الجامعة فالرجاء أن تخاطبني باللغة العربية الفصحى.	ASU provides many of specialists in different fields and this system is designed to facilitate your communications with the university, so please speak with me using modern Arabic.				
4	متى يبدأ التسجيل في الجامعة	When is the registration will start in the university	الحد الأدنى لمعدل القبول والتسجيل في التخصصات العلمية هو 80% في الثانوية العامة من الفرع العلمي	The minimum rate of acceptance and registration in scientific disciplines is 80% in the high school section of scientific	11,33			

According to Figure 1, the “ArabChat Evaluator” performs the following steps for each unevaluated conversation in the “Brief Log”:

- 1) Read the utterance and the response using the “Utterance and Response Reader” module.
- 2) Extracts the “Information Requirements” from the utterance using “Extracted Information Requirement” module
- 3) Retrieves the “Information Requirements” for the response using the “Retrieve Information Requirements” module
- 4) Undertakes a comparison between the two generated “Information Requirements” lists and then take its decision (evaluating a conversation) using the “Compare Lists Make Decision” module

After evaluating the first unevaluated conversation in the “Brief Log”, the “ArabChat Evaluator” checks the entries of the “Brief Log”. If unevaluated records still exist, it starts reading and repeats the previous steps until it is finish checking all the log’s records and then it evaluates the ArabChat. The following components for the “ArabChat Evaluator” framework will now be discussed:

- **The “Utterance and Response Reader” module:** The “ArabChat Evaluator” starts reading the utterance and the response using the “Utterance and Response

Reader” module from the “Brief Log” located in ArabChat’s brain. Then, in order to retrieve its “Information Requirements”, the “ArabChat Evaluator” sends the utterance and the response to the “Extracted Information Requirement” and the “Retrieve Information Requirements” module respectively.

- **The “Extracted Information Requirement” module:** extracts the “Information Requirements” from the utterance by tokenising the utterance and converting it into a list of words. Then a stemming process is done on the list of words that converts it to a stemmed list of words, called the “Utter List”. The stemming process is based on an Arabic based stemming algorithm with a good performance and it explained in [28]. Then, the “Utter List” elements are matched with ArabChat “Information Requirements” list elements. If a matching occurs, the group number of the matched word in the ArabChat “Information Requirements” list is replaced by the matched word in “Utter List”. Otherwise, the unmatched word in the “Utter List” is removed. Finally, the “Utter List” contains only numbers that represent the groups matched to the utterance’s stemmed words. Eventually, “ArabChat Evaluator” removes the duplication of the same number from the “Utter List”, if they exist.

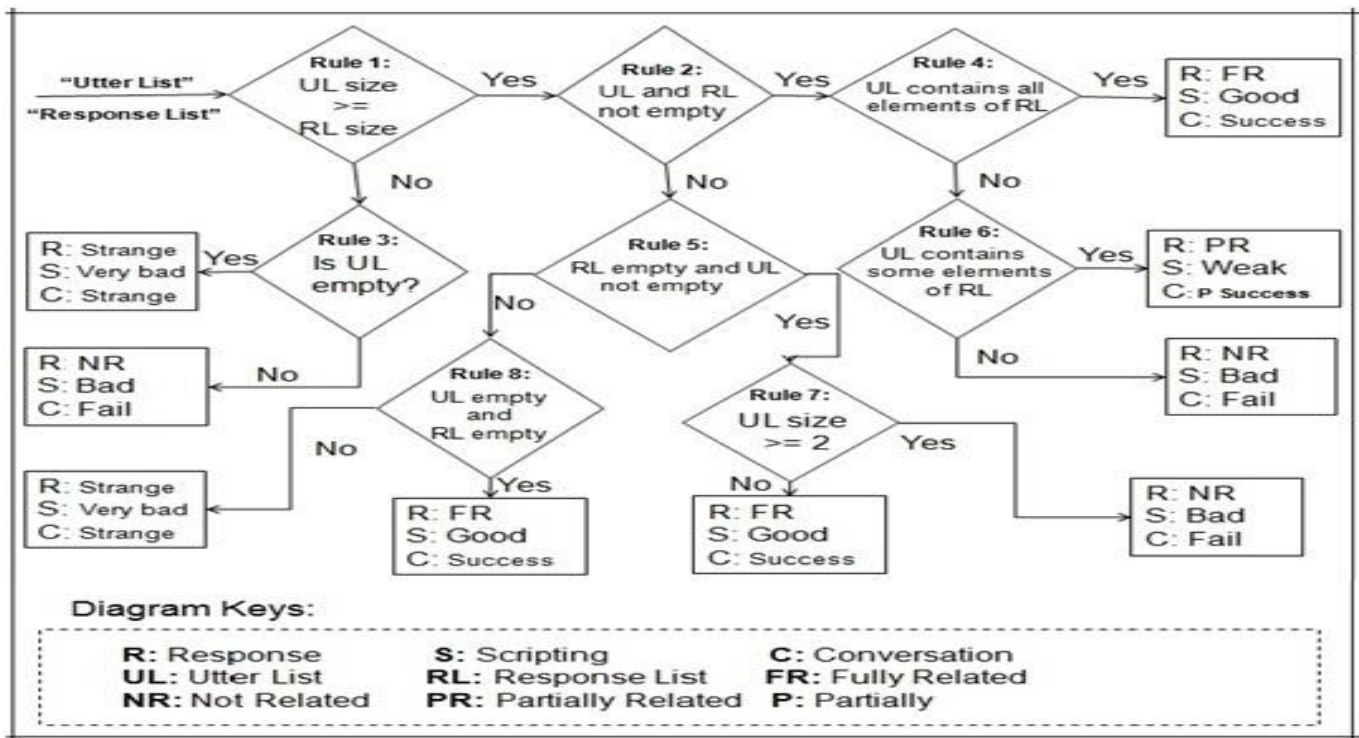


Fig. 2. The “Compare Lists Make Decision” module methodology

- **The “Retrieve Information Requirements” module:** retrieves the “Information Requirements” for the rule that the processed response belongs to. However, the “Information Requirements” is already stored as a collection of numbers in the “Brief Log” that was accumulated by ArabChat during users’ conversations. Then, it starts removing the duplicated numbers that might be caused by the accumulative process and calls it the “Response List”.
- **“Compare Lists Make Decision”:** both of the generated lists (“Utter List” and “Response List”) are sent to the “Lists Comparing Make Decision” module in order to start the comparison and evaluation of the processed conversation. Figure 2 represents the “Compare Lists Make Decision” module methodology.

The “ArabChat Evaluator” evaluates three aspects of each conversation, “Response”, “Scripting”, and “Conversation”. For each aspect, different results might be generated as presented in Table 3. In this table, each evaluation aspect has 4 potential results. For instance, the “Response” aspect has 4 results which are either “Strange”, “Not related”, “Partially related” or “Fully related”.

The “Response” evaluation aspect’s result will determine the results of the rest aspects (the “Scripting” and “Conversation”). For instance, if the “Response” result is “Not related”, then the “Scripting” results will be “Bad” and the “Conversation” results “Fail”. More detailed explanations for these evaluation aspects and the differences between their results will be described later in this section when the “Compare Lists Make Decision” module methodology is described.

TABLE III. EVALUATION ASPECTS RESULTS

Evaluation aspect	Evaluation result
Response	Strange
	Not related
	Partially related
	Fully related
Scripting	Very bad
	Bad
	Weak
	Good
Conversation	Strange
	Fail
	Partially success
	Success

The following rules that presented in Figure 2 explain the methodology of the “Compare Lists Make Decision” module:

**Rule 1:** Check the “Utter List” and “Response List” sizes. If “Utter List” size  $\geq$  “Response List” size, go to Rule 2; otherwise go to Rule 3.

**Rule 1 description:** rule 1 is the first rule that deals with the two lists and is responsible for checking their sizes. When the “Utter List” is not empty, it means the user entered at least one keyword that already matched one of the applied domain’s keywords. When the “Response List” is not empty, it means that the ArabChat fired a rule after a matching occurred between the user’s utterance and a pattern. Default rules have no “Information Requirements” values; i.e. if a default rule is fired, the “Response List” will be empty.

When the size of the “Utter List” is greater than or equal to the size of the “Response List”, it concludes on one of two meanings: firstly, both lists have elements and that the “Utter List” elements are greater than the “Response List” elements, which means that the user has entered a greater amount of keywords than required to fire the processed a rule (the fired rule that has the processed response). Secondly, both lists are empty (their sizes are zeros) and consequently are equalled.

**Rule 2:** Check the “Utter List” and “Response List” contents. If none of them are empty, continue to Rule 4; otherwise continue to Rule 5.

**Rule 2 description:** Rule 2 comes after Rule 1 in two conditions; either both lists are empty or the “Utter List” size is greater than the “Response List” size. Therefore, Rule 2 is responsible for limiting these probabilities by checking if the two lists are not empty, if the two lists are not empty, a user entered at least one keyword and ArabChat fired a rule other than the default rule.

**Rule 3:** Check the “Utter List” contents. If empty, then the evaluations for the three aspects are:

**Response:** Strange.

**Patterns Scripting:** Very bad.

**Conversation:** Strange.

If the “Utter List” is not empty, then the evaluations for the three aspects are:

**Response:** Not related.

**Patterns Scripting:** Bad

**Conversation:** Fail.

**Rule 3 description:** Rule 3 comes after Rule 1 confirms that “Response List” size is greater than “Utter List” size. However, Rule 3 has two probabilities; either the “Utter List” size is smaller than “Response List” size or the “Utter List” is empty. Therefore, Rule 3 checks the “Utter List” size, whether it is empty or not. If empty, it means that a user has not entered any keyword related to the applied domain. Therefore, the “ArabChat Evaluator” will evaluate the conversation as “Strange”, because if the user entered no keywords, the ArabChat should fire a default rule and then the “Response List” should be empty. As a result, the response evaluated was “Strange”, thus indicating that the patterns scripting for the targeted rule is “Very bad”. In contrast, if the “Utter List” is not empty, it means that a user has entered at least one keyword and the “Information Requirements” of the fired rule, whether default or not, is less than what the user entered. Therefore, the “ArabChat Evaluator” decided to evaluate the

conversation as “Fail” because the response is “Not related” and thus the patterns scripting is “Bad”.

**Rule 4:** Check the “Utter List” and “Response List” contents. If all “Response List” elements are in “Utter List” list, then the evaluations for the three aspects are:

**Response:** Fully related

**Patterns Scripting:** Good

**Conversation:** Success.

Otherwise (not all “Response List” elements are in “Utter List” list), continue to Rule 6.

**Rule 4 description:** Rule 4 comes after Rule 2 if it is agreed that the two lists are not empty. Therefore, the “ArabChat Evaluator” tests through Rule 4 if all elements of the “Response List” are in the “Utter List” list. If so, the ArabChat fires a rule that meets all the utterance requirements of the user, and thus, the response is evaluated as “Fully related” and the conversation is evaluated as “Success”. Consequently, the pattern scripting for the fired rule is “Good”. In contrast, if not all elements of the “Response List” are in the “Utter List”, the “ArabChat Evaluator” will continue to Rule 6, which is responsible for testing if some elements of “Response List” are in the “Utter List”.

**Rule 5:** Check the “Utter List” and “Response List” contents. If the “Utter List” is not empty and “Response List” is empty, continue to rule 7; otherwise continue to Rule 8.

**Rule 5 description:** Rule 5 comes after Rule 2 on the condition that at least one of the two lists is empty. Therefore, Rules 5 is used to test if the “Utter List” is not empty and the “Response List” is empty. If so, this means that the ArabChat fired a default rule but a user had entered at least one keyword. Consequently, the “ArabChat Evaluator” continues to Rule 7 to check if the “Utter List” size is greater than or equal to 2 (Threshold), which means the user entered enough “Information Requirements” in his/her utterance. Otherwise, “ArabChat Evaluator” will continue to Rule 8 which is responsible to test if both lists are empty. Experimentally, it was determined that the minimum number of keywords that should be in a matched utterance is 2, which is considered a threshold point.

**Rule 6:** Check the “Utter List” and “Response List” contents. If some “Response List” elements are in the “Utter List”, then the evaluations for the three aspects are:

**Response:** Partially related.

**Patterns Scripting:** Weak.

**Conversation:** Partial success.

Otherwise (no element of “Response List” is in “Utter List” list), then the evaluations for the three aspects are:

**Response:** Not related.

**Patterns Scripting:** Bad.

**Conversation:** Fail.

**Rule 6 description:** Rule 6 comes after Rule 4 on the condition that not all “Response List” elements are in the “Utter List”. Therefore, Rule 6 is used if some (one or more but not all) of the “Response List” elements are in the “Utter List”. If so, a user has entered keywords in his/her utterance and ArabChat fired a rule that met some of the user’s utterance requirements. Thus, the “ArabChat Evaluator” evaluates the response as “Partially related” and the conversation as “Partial success” because the fired rule replied to some of the user’s requirements but not all of them. Consequently, this indicates that the patterns scripting for the fired rule is “Weak”. Otherwise, the user entered keywords and ArabChat fired a rule not related to the user’s utterance requirement at all. Therefore, the “ArabChat Evaluator” evaluates the response as “Not related” and the conversation as “Fail”, as the scripting of the fired rule pattern is “Bad”.

**Rule 7:** Check the “Utter List” size. If “Utter List” size  $\geq 2$ , then the evaluations for the three aspects are:

**Response:** Not related

**Patterns Scripting:** Bad

**Conversation:** Fail.

Otherwise (“Utter List” size  $< 2$ ), then the evaluations for the three aspects are:

**Response:** Fully related.

**Patterns Scripting:** Good.

**Conversation:** Success.

**Rule 7 description:** Rule 7 comes after Rule 5 on the condition that the “Utter List” is not empty while the “Response List” is empty. Therefore, Rule 7 tests if the “Utter List” size is greater than, or equal to 2 (the threshold point). If so, this means a user entered a minimum of 2 keywords, and ArabChat fired a default rule because the “Response List” was empty. The “ArabChat Evaluator” evaluates the response as “Not related” and the conversation as “Fail”, because the scripting of the fired rule patterns is “Bad”. Otherwise, the user enters less than 2 keywords, which is below the determined threshold point. Therefore, the “ArabChat Evaluator” evaluates the response as “Fully related” and the conversation as “Success”, thus the pattern scripting is “good” as long as the utterance is outside the scripted domain. Entering an amount of keywords less than the threshold with an empty “Response List” means that a user entered an utterance outside the applied domain and ArabChat is only responsible to reply to utterances inside the applied domain, thus the conversation is considered successful.

**Rule 8:** Check the “Utter List” and “Response List” contents. If both lists are empty, then the evaluations for the three aspects are:

**Response:** Fully related.

**Patterns Scripting:** Good.

**Conversation:** Success.

Otherwise, then the evaluations for the three aspects are:

**Response:** Strange.

**Patterns Scripting:** Very bad.

**Conversation:** Strange.

**Rule 8 description:** Rule 8 comes after Rule 5 on the condition that the “Utter List” is not empty while the “Response List” is empty. Therefore, Rule 8 is used if both lists are empty. If so, this means that a user entered an utterance outside the applied domain, and ArabChat fired a default rule. Thus, the “ArabChat Evaluator” evaluates the response as “Fully related” and the conversation as “Success” and patterns scripting as “Good”. Otherwise, the “Response List” was not empty, while the “Utter List” was empty. In other words, a user entered no keywords and the ArabChat fired a rule that requires keywords in the processed utterance to be fired. However, it might be impossible for this case to happen. If it does, there is something strange in patterns scripting or in the rule “Information Requirements” list. Therefore, “ArabChat Evaluator” evaluates the response and the conversation as “Strange” and thus, the patterns scripting result is “Very bad”.

As the “ArabChat Evaluator” is mainly used to evaluate the quality of response generated by the ArabChat scripting engine, it may be a good indicator of the whole ArabChat performance. When a response is evaluated as “Fully related”, it means that the whole utterance’s “Information Requirements” are replied. Consequently, this means that the scripting engine fires the best rule, indicating that the scripting engine methodology worked properly. On the other hand, firing the best rule indicates that the rule scripting in general, and especially its patterns scripting, are scripted well. Therefore, the scripting is evaluated as “good”, and the conversation is evaluated as “Success”. However, the “Fully related” response might be caused by the user entering an utterance outside the applied domain. ArabChat is only responsible to reply utterances related to the applied domain. Regardless of the reason of evaluating a response as “Fully related”, it is obvious that not only the response is evaluated. Implicitly, ArabChat’s scripting engine and the whole ArabChat is evaluated by this process.

By applying the “ArabChat Evaluator” to the utterances presented in Table 2, the evaluation results are achieved and presented in Table 4. The “ArabChat Evaluator” can be run at any time in order to evaluate the unevaluated utterances in the “Brief Log”. Each time the “ArabChat Evaluator” runs, it checks that the evaluation fields (response, pattern scripting, and conversation) of the utterances are empty and starts evaluating them.

As discussed earlier, the “ArabChat evaluator” aims to evaluate the ArabChat in terms of its generated quality of response. Thus, it acts as an indicator of ArabChat’s scripting engine performance and reflects the quality of pattern scripting. Subsequently, user satisfaction with ArabChat can be determined using the same methodology (ArabChat evaluator).

TABLE IV. A CUSTOMISED SAMPLE OF THE “BRIEF LOG” COMPONENTS AFTER THE EVALUATION PROCESS

#	Utterance	Utterance in English	Response in Arabic	Response in English	Information Requirement	Response evaluation	Scripting evaluation	Conversation evaluation
1	كم سعر ساعه تخصص الحاسوب	What is the price of a credit hour of computing course.	سعر الساعة المعتمدة لتخصص علم الحاسوب هو 75 دينار أردني	The price of a credit hour for computing course is 75 JD which is equaled to 106 USD	8,8,48	Fully related	Good	Success
2	كم عدد الساعات لتخصص التمريض والمحاسبة	What is the price of a credit hour of nursing accounting.	عدد الساعات المعتمده لتخصص علم الحاسوب هو 132 ساعة معتمدة	Total number of credit hours of Computer Science is a 132 hours.	29,29	Partially related	Weak	Partially success
3	لقد سمعت من أصدقائي بأنك ثرثار	I have heard from my friends that you are talkative	تقدم جامعة العلوم التطبيقية العديد من التخصصات في شتى المجالات وقد صمم هذا النظام ليقيم بالتسهيل عليك عملية الاستفسار والدرشة مع موظفي الجامعة فالرجاء أن تخاطبني باللغة العربية الفصحى.	ASU provides many of specialists in different fields and this system is designed to facilitate your communications with the university, so please speak with me using modern Arabic.		Fully related	Good	Success
4	متى يبدأ التسجيل في الجامعة	When is the registration will start in the university	الحد الأدنى لمعدل القبول والتسجيل في التخصصات العلمية هو 80% في الثانوية العامة من الفرع العلمي	The minimum rate of acceptance and registration in scientific disciplines is 80% in the high school section of scientific	11,33	Not related	Bad	Fail

The “ArabChat evaluator” was run in order to read the reported records from the “Brief Log” and fill the empty fields, which are “Response Evaluation”, “Scripting Evaluation”, and “Conversation Evaluation”. Each record represents one conversation (utterance) between a user and ArabChat. The “ArabChat evaluator” analyses these conversations one by one as described above. Users’ conversations are categorised by the user sequence numbers. Therefore, it is possible to determine the satisfaction per user and the average rate of general user satisfaction. The user satisfaction is determined based on the evaluated parameter (Conversation Evaluation), which indicates the status of a user conversation with ArabChat. However, determining the average user satisfaction through the descriptive evaluated results (“Success”, “Partially success”, “Fail” and “Strange”) is not an easy task. Therefore, a new technique was developed, called the “ArabChat Evaluation Calculator”, which converts these descriptive evaluated results into numeric values and calculates the user satisfaction value.

The “ArabChat Evaluation Calculator” is based on the number of utterances per user in order to assign a numeric value for the evaluated results (“Success”, “Partially Success”, “Fail” and “Strange”). In this technique (the “ArabChat Evaluation Calculator”), the “Strange”

conversation will be considered as “Fail” conversation. Equations 1, 2 and 3 are used to assign the equivalent numeric values for the evaluated results:

$$NVSC = \frac{100}{\text{Num.Utters}} \quad (1)$$

$$NVPSC = \frac{NVSC}{2} \quad (2)$$

$$NVFC = 0 \quad (3)$$

Where,

NVSC: Numeric Value Success Conversation.

NVPSC: Numeric Value Partially Success Conversation.

NVFC: Numeric Value Fail Conversation.

Num. Utters: Number of utterances.

Equation 1 represents the converting mechanism for the evaluated result “Success” into its equivalent numeric value by assuming that the user satisfaction is 100%. Therefore, a division process has been performed, as Equation 1 showed, by taking into consideration the total number of utterances for the evaluated user. For instance, if a user entered 5 utterances, the “ArabChat Evaluation Calculator” calculates the NVSC as 100/5 and thus NVSC=20.

Then, the “ArabChat Evaluation Calculator” applies Equation 2 in order to calculate the NVPC according to the equation, as  $20/2$  and thus,  $NVPC=10$ . Finally, the NVFC is calculated according to Equation 3, which always assigns a zero value to the NVFC for “Fail” and “Strange” conversations.

Naturally, user satisfaction might improve or contract during prolonged use of ArabChat. Therefore, determining the user satisfaction by calculating the average of the numeric evaluated results might not be quite accurate. Instead, an exponential average calculation method has been applied on the numeric results in order to calculate a more natural result. The exponential average calculation method is based on the consecutive conversations of the evaluated results for the same result type, such as consecutive “success”, “partially success” or “fail conversations”. The following examples show how the “ArabChat Evaluation Calculator” based on the exponential average calculation assigns the equivalent numeric values for the consecutively evaluated results:

The First success conversation = NVSC.

The Second consecutive success conversation =  $NVSC^{1.01}$ .

The Third consecutive success conversation =  $NVSC^{1.02}$

The Fourth consecutive success conversation =  $NVSC^{1.03}$

The Eleventh consecutive success conversation =  $NVSC^{1.10}$

And so on.

The First partially success conversation = NVPC.

The Second consecutive partially success conversation =  $NVPC^{1.01}$ .

The Third consecutive partially success conversation =  $NVPC^{1.02}$

The Fourth consecutive partially success conversation =  $NVPC^{1.03}$

The Eleventh consecutive partially success conversation =  $NVPC^{1.10}$

And so on.

The First fail conversation = NVFC.

The Second consecutive fail conversation = NVFC.

The Third consecutive fail conversation = NVFC.

The Fourth consecutive fail conversation = NVFC.

The Eleventh consecutive fail conversation = NVFC

And so on.

The next section discusses the evaluation results of the “Black box approach” after applying the “The ArabChat Evaluator” on the conversations of ArabChat experiment’s users.

#### IV. THE EVALUATION

The evaluation for the proposed and developed methodology has been done through conducting an real experiment on the selected CA(ArabChat). The experiment was conducted to test the full ArabChat capabilities from different aspects through applying the developed methodology (ArabChat Evaluator). As mentioned before, this proposed and developed framework will be able to test the ArabChat engine in terms of its ability to match utterances properly and it can test the user satisfaction in general.

ArabChat was deployed on the ASU website [29] and accessed by all qualified users such as registered students, non registered students, and employees. ArabChat was available online and in use for 23 days.

The ArabChat handled 1766 utterances from 203 users, an average of 8.699 utterances per user.

#### Evaluation results and discussion

After applying the “ArabChat Evaluator” and the “ArabChat Evaluation Calculator” on the ArabChat users’ conversations in the experiment (203 users with 1766 utterances), the average of ArabChat users’ satisfaction is = 64.31%.

A manual checking for all users’ conversations has been done in order to classify them as serious users and unserious users. The serious user who keep talking inside the selected applied domain (information point advisor for ASU). Where the unserious user who just try to trick the ArabChat, saying something funny or his/her utterances has impolite words. This manual checking has been raised that an 8.267% of users’ conversations were placed in the second category which might reveal the existence of unserious users who negatively affected the evaluation result (users’ satisfaction is 64.31%). As a result, this outcome (64.31%) could be considered as a reasonable result of the average of user satisfaction.

#### V. SUMMARY

The fundamental purpose behind a system’s evaluation is to improve its performance. As discussed in this paper, the lack of a comprehensive evaluation framework has been a limiting factor in the growth of Conversational Agents. In addition, different types of CAs might require different frameworks of evaluation.

Furthermore, devising an automatic method for evaluating CAs is not an easy task as a user utterance might have a rich semantic meaning, which is hard to automatically detect. In addition, CA conversations vary among users even for closed applied domains.

Generally, chatting with a CA does not mean that a user will keep entering either questions or non-questions only. The natural conversations between a user and a CA should consist of both (questions and non-questions). Nevertheless, the amount of question and non-question utterances might be based on the following factors:

1) The topical nature of a CA's applied domain; for instance, an entertainment domain might differ from an information point advisor.

2) The users, if they are familiar with the nature of a CA. It can be concluded from experiment 1 of ArabChat that many users consider ArabChat a question answering system. As a result, a large amount of questions were entered. In addition, 92.3% of experiment 1's users confirmed that they had never used any similar service before, which points to a lack of experience in handling these services.

3) The way a CA forms its response might also encourage a user to ask questions or continue chatting with non-question utterances.

The ArabChat was evaluated depending on the ratio of matched and unmatched utterances [1]. This technique might give a general overview about the ArabChat's performance. However, it cannot give an accurate result as the utterances might matching wrong rules. Therefore, in this paper a new framework has been modelled and developed to evaluate the ArabChat in a more accurate way. By the proposed framework, the evaluation focused on the "Information Requirements" that should be shared between the utterance and the fired rule. According to the conducted experiment and the evaluation of the ArabChat based on the proposed framework (ArabChat Evaluator and ArabChat Calculator) and based on the experiment's results, it can be concluded that ArabChat successfully handled conversations for ASU students.

#### ACKNOWLEDGMENT

The authors are grateful to the Applied Science Private University, Amman, Jordan, for the full financial support granted to this research.

#### REFERENCES

- [1] Hijjawi, M., et al. ArabChat: An Arabic Conversational Agent. in proceeding of the 6th International Conference on Computer Science and Information Technology (CSIT). 2014. Amman, Jordan: IEEE Explore.
- [2] Turing, A., Computing machinery and intelligence. Mind, 1950: p. 433-460.
- [3] Hijjawi, M., ArabChat: An Arabic Conversational Agent. PhD thesis, in School of Computing. 2011, Manchester Metropolitan University: Manchester. p. 241.
- [4] Goh, O., A framework and evaluation of Conversational Agents. PhD thesis, in Information Technology school. 2008, Murdoch University.
- [5] Russell, s. and P. Norvig, Artificial Intelligence A Modern Approach. . Vol. Third edition. . 2010: Pearson Education.
- [6] Sammut, C. and D. Michie, InfochatTM Scriptor's Manual, Convagent Ltd. . 2001: Manchester.
- [7] Weizenbaum, J., ELIZA-A computer program for the study of natural language communication between man and machine. Communications of the ACM., 1966. Vol 10.: p. PP 36-45.
- [8] Wallace, R. ALICE: Artificial Intelligence Foundation Inc. . 2008 [cited; Available from: <http://www.alicebot.org>.
- [9] Crystal, D., Dictionary of linguistics and phonetics., Blackwell., Editor. 2008.
- [10] Habash, N., Introduction to Arabic Natural Language Processing, ed. U.o.T. Graeme Hirst. 2010: Morgan & Claypool.
- [11] O'Shea, K., Z. Bandar, and K. Crockett, A Novel Approach for Constructing Conversational Agents using Sentence Similarity Measures. 2008.
- [12] Almarsoomi, F., et al., Arabic Word Semantic Similarity. International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering. World Academy of Science, Engineering and Technology, 2012. Vol:6, No:10.
- [13] David, D., S. Karen, and J. rck, Natural language processing for information retrieval. . Commun. ACM., 1996. Volume 39: p. pp. 92-101.
- [14] Walker, M., et al. PARADISE: A framework for evaluating spoken dialogue agents. In Proc. of 35th ACL. 1997.
- [15] Sanders, G. and J. Scholtz, Measurement and evaluation of embodied conversational agents. MIT Press, 2000.
- [16] Hugh, L. Loebner's Home Page. Available at: [www.loebner.net](http://www.loebner.net). 2016 [cited Accessed in 2016].
- [17] Maier, E., M. Mast, and S. LuperFoy, Dialogue Processing in Spoken Language Systems., in ECAT96 Workshop. 1996: Hungary.
- [18] Hirschman, L., The Roles of language processing in a spoken language interface. National Academy Press Washinton- Voice Communication Between Humans and Machines, 1995: p. pp217-37.
- [19] Simpson, A. and F. Norman. Black box and glass box evaluation of the SUNDIAL system. in Proceedings of the 3rd European Conference on Speech Communication and Technology. 1993.
- [20] ANDERSEN, V., et al., A methodological approach for designing and evaluating intelligent applications for digital collections. . Applied Artificial Intelligence., 2003. Vol 17(Issue 8-9).
- [21] Lamel, L., Bennacef, S., Gauvain, J. L., Dartigues, H. and Temem, J. N. , User evaluation of the MASK kiosk. Speech Commun., 2002. Vol 38(1): p. PP 131-39.
- [22] Xiang, Y. and C. Yam, Design and evaluation of Elva: an embodied tour guide in an interactive virtual art gallery. Research Articles. Comput. Animat. Virtual Worlds., 2005. Vol 16(2): p. pp. 109-19.
- [23] Goh, O., et al., A Black-box Approach for Response Quality Evaluation of Conversational Agent Systems. . World Academy of Science Engineering and Technology., 2007. Vol 3: p. PP 195-203.
- [24] Kopp, S., et al., A conversational agent as museum guide: design and evaluation of a real-world application. Lecture Notes in Computer Science. ed., Springer-Verlag, 2005.
- [25] Le Bigot, L., E. Jamet, and J. Rouet, Searching information with a natural language dialogue system: a comparison of spoken vs. written modalities. Applied Ergonomics, 2004. Vol 35(6): p. pp. 557-64.
- [26] Abu Shawar, B. and E. Atwell. Different measurements metrics to evaluate a chatbot system. . in Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies, Rochester, New York: Association for Computational Linguistics. 2007.
- [27] Lynette, H. and T. Henry, Overview of evaluation in speech and natural language processing. IN Survey of the state of the art in human language technology. ed., Cambridge University Press, 1997.
- [28] Hijjawi, M., et al. An Application of Pattern Matching Stemmer in Arabic Dialogue System. . in 5th International KES Conference on Agents and Multi-agent Systems – Technologies and Applications. 2011. Manchester, UK: Elsiver.
- [29] ASU. Applied Science University- ArabChat. 2011 [cited 2011; Available from: Available at: [www.asu.edu.jo](http://www.asu.edu.jo).

# Effects of Walls and Floors in Indoor Localization Using Tracking Algorithm

Farhat M. A. Zargoun

Computer and Systems Engineering  
department  
Faculty of Engineering, Zagazig  
University, Egypt

Ibrahim M. Henawy

Computer Science department  
Faculty of computer and information,  
Zagazig University, Egypt

Nesreen I. Ziedan

Computer and Systems Engineering  
department  
Faculty of Engineering, Zagazig  
University, Egypt

**Abstract**—The advancement in wireless and mobile networks has led to an increase in location based services (LBS). LBS can be applied in many applications, such as vehicle systems, security systems, and patient tracking systems.

The Global Navigation Satellite Systems (GNSS) have become very popular due to their ability to provide highly accurate positions, especially in outdoor environments. However, GNSS signals become very weak when they go through natural or man-made structure, like in urban canyons or indoor environments. This hinders the applicability of GNSS-based localization techniques in such challenging environments.

Many indoor localization techniques are based on the received signal strength (RSS). An RSS is proportional to the distance to an access point (AP), where it is stronger in power when it is closer to an AP, given that the received signal is not obstructed by walls or floors. This paper aims at studying the effect of walls and floors on the RSS, and estimating the distribution of the RSS due to such obstructions. Moreover, a tracking algorithm based on a multi-walls and floors propagation model is applied to increase the positioning accuracy.

**Keywords**—Indoor localization; Tracking algorithm; Effects of wall and floor on RSS; Effects of obstruction; Multi wall and floor propagation model

## I. INTRODUCTION

The received signal strength (RSS) from a transmitter to a receiver is proportional to the distance between them. Therefore, the RSS has been an important measure in wireless positioning techniques. Three transmitters, or access points (AP), are required to estimate the 3-D position of a receiver.

The study provided in this paper takes into account walls and floors as obstructions that attenuate the RSS. The losses in the RSS depend on materials that makeup walls and floors, signals can go through walls or floors, but these signals suffer from attenuation that depends on the type and thickness of the walls and floors. RSS in indoor locations is difficult to predict due to the attenuation effects. However, signal losses can be measured using a multi walls and floors propagation model.

The remainder of this paper is organized as follows. First, the path loss and attenuation due to walls and floors are introduced. Following that, an experiment that illustrates the distribution of RSS is presented. A method is then presented to predict the location using RSS. The effect of multi-walls and multi-floors on the RSS is investigated next, and used in an algorithm to estimate the location in an indoor environment.

Finally, experiments are performed to investigate the validity of the presented algorithm.

## II. RELATED WORK

In [4] K-Nearest Neighbor (KNN) algorithm is used to predict the user's location in a single floor environment. KNN works as a matching technique that compares a measured RSS with RSS stored in a radio map, and concludes the position as the one that has stored RSS closest to the measured RSS. The matching can be done by using metric techniques such as the Euclidean distance [4]. The estimated location accuracy is affected by the size of reference points and k-value.

The authors in [5] designed an indoor localization system by using Bayesian graphical model. This model depends on the set of conditional independence relationships. The first model of Bayesian network is used by [6], which was a simple model to estimate the user's location for a single floor. Sampling techniques, which are needed for the Bayesian model, are calculated by using collected data. One of the most important sample techniques is based on the Markov Chain Monte Carlo (MCMC) [6].

In indoor environments, walls and floors attenuate the RSS and hinder the position estimation.

According to [7][9], the signal strength is affected by the types of concrete walls. Some types of walls have a high attenuation factor, so that each floor can reduce the received signal strength between 10 and 35 dBm. This factor plays a main role in estimating a user's location in a multi-floor environment. There are some propagation mechanisms that need to be determined in order to investigate the characteristics of propagation in a multi-floor environment, such as multipath diffraction, reflection, and scattering. The diffraction or scattering due to a window frame or an edge of a building wall can also affect the FAF. Moreover, the attenuation of the RSS depends on the size of the window frame in each floor.

## III. PATH LOSS AND ATTENUATION

A path loss is a measure of the average signal attenuation. It can be expressed as:

$$P_L = P_{L0} + 10n \log(d) \quad (1)$$

Where

$P_L$  - Attenuation due to multi walls and floors.



$PL_0$  - Path loss at distance of 1 meter.

$n$  - Power-law relationship between distance and received power.

$d$  - Distance between a transmitter and a receiver.

In order to take into account the attenuation due to walls and floors, two additional terms are added to (1). The new equation is called multi walls and floors propagation equation [1]. The relationship between penetration loss and the number/type of floors and walls can be expressed by:

$$PL_{mwf} = PL_0 + 10n \log(d) + \sum_{i=1}^I \sum_{k=1}^{K_{wi}} PL_{wik} + \sum_{j=1}^J \sum_{k=1}^{K_{fj}} PL_{fjk} \quad (2)$$

$PL_{mwf}$  - Attenuation due to multi walls and floors.

$PL_0$  - Path loss at distance of 1 meter.

$n$  - Power-law relationship between distance and received power.

$d$  - Distance between transmitter and receiver.

$PL_{wik}$  - Attenuation due to wall type I and Kth traversed walls.

$PL_{fjk}$  - Attenuation due to floor type j and Kth traversed floors.

$i$  - Number of walls.

$j$  - Number of floors.

$K_{wi}$  - Number of traversed walls.

$K_{fj}$  - Number of traversed floors.

#### IV. DISTRIBUTION OF RSS

To design an indoor localization system, one needs to study the properties of the RSS. RSS is difficult to predict due to factors such as multipath effects and obstructions, Figures 2(a) and 2(b) show the measured RSS to 2 access points (APs). The experiment was performed in the first floor in the Training Center (TC), in the faculty of electronic engineering, Bani-Walid, Libya. The walls are made of wood that has iron bars inside. The RSS were measured at fixed locations for 50 minutes, with 10 seconds interval, during work hours. The figures show that each experiment is different from each the other. This is because of the difference in the distance between the sender and the receiver, and because of obstructions.

Two experiments are conducted to test the difference between normal log of signal propagation.

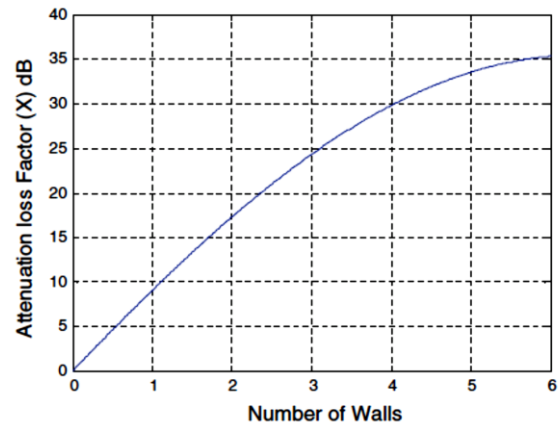


Fig. 1. Best fit curve of the average curves of the attenuation loss factor

One of them when there is a line of sight (LOS) between the transmitter and the receiver, and another one show how the obstruction can affect the signal [13].

The experiment setup is two access points (reference nodes) and one laptop computer (user node), Xirrus WiFi inspector software is installed in user node to capture the signal strength. The RSS is measured from two different distances from each AP, in the presence of obstructions such as walls and floors. Another experiment is conducted to measure the RSS without obstruction. Figures 3(a), 3(b) show the results.

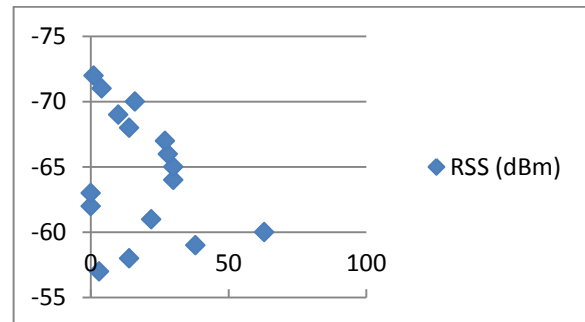


Fig. 2. (a) AP 1, LOS

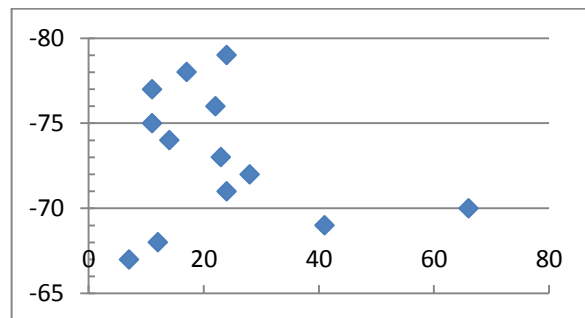


Fig. 2. (b) AP 1, NLOS

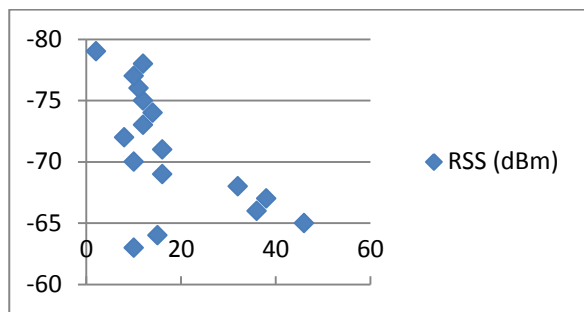


Fig. 3. (a) AP 2, LOS

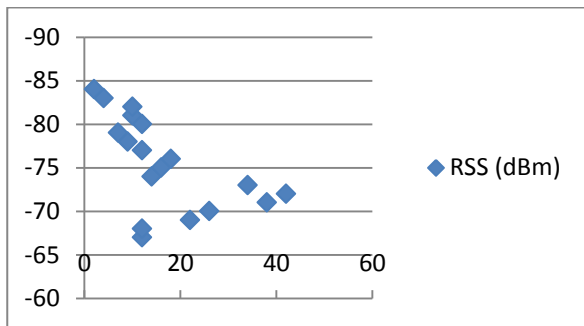


Fig. 3. (b) AP 2, NLOS Histogram of RSS for two access points

### V. USING RSS TO PREDICT THE LOCATION

RSS can be used as a reference to predict the indoor location. A test is conducted to show the RSS variability due to distance and obstruction. Figure (3) shows the variation of the RSS measurements recorded from four APs, where two of them are located in the first floor, and the others are located in the second floor [13]. The measurements are recorded by a person walking through a track in the first floor at the Training Center Building. The signal received at any given location is stronger when the location is close to the AP and weaker when it is far away. This shows the feasibility of using RSS as a location Fingerprint. Figure (5) shows the uniqueness of the RSS tuples. Each RSS tuples at each location are different. This indicates that RSS fingerprints are a reasonable choice for inferring indoor locations. Figure (5) also shows that there is a small variation in the RSS when the distance between two training points is short.

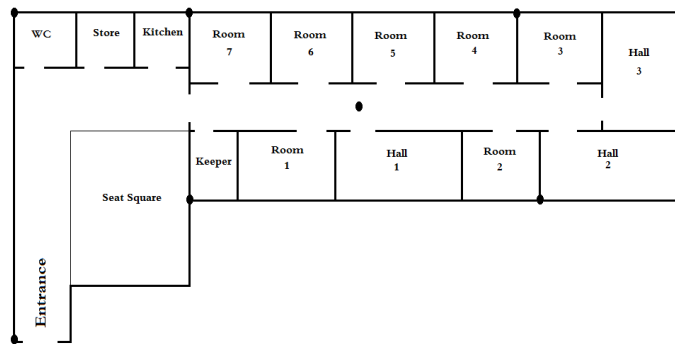


Fig. 4. Training Center Building

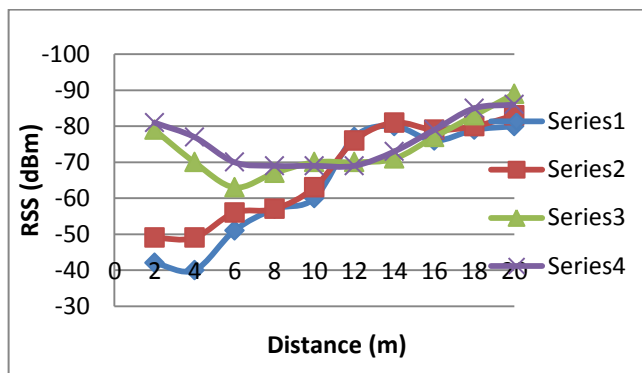


Fig. 5. RSS while walking

Table (1) shows the RSS statistics from the four APs at two different floor

TABLE I. RSS STATISTICS

Statistics	AP 1 1st floor	AP 2 1st floor	AP 3 2nd floor	AP 4 2nd floor
Mean Measured	-64.2	-67.3	-73.9	-75.8
Mean Calculated	-61.8	-63.4	-68.1	-71.2
Standard Deviation	1.2	1.95	2.9	2.3
Min.	-40	-49	-63	-69
Max.	-80	-83	-89	-86

### VI. MULTI WALL EFFECT

According to [1], the reduction in RSS depends on type of the wall. Table (2) shows the attenuation in the RSS due to different types of walls.

TABLE II. DIFFERENT TYPES OF WALLS ATTENUATION

Wall type and thickness	First wall	Second wall
Concrete 10 cm	16 db	14 db
Concrete 20 cm	29 db	24 db

An experiment is performed in another building, to show the effect of concrete walls. The walls have 20cm thickness. 3 offices are used, where each office has one AP. The signal strength is measured at the middle of each office, where the measurements are done over 5 minutes with 3 seconds intervals, Figure(7) shows the result of this experiments.

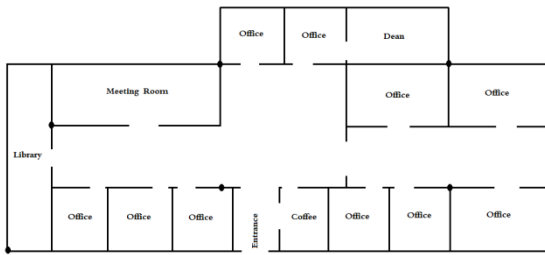


Fig. 6. Administration Building

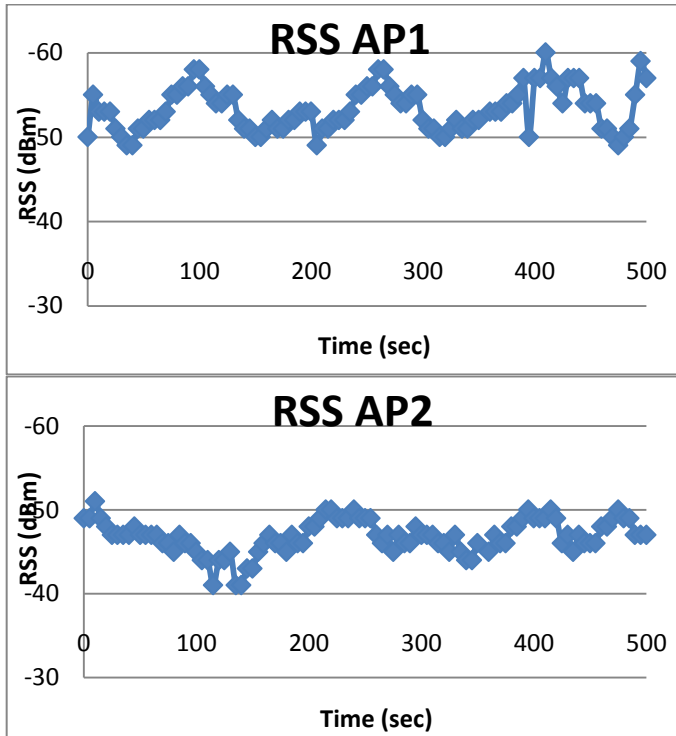


Fig. 7. Effects of walls

VII. MULTI FLOOR EFFECT

According to [8], table (3) shows the attenuation in RSS caused by floors. Another experiment is conducted to check the attenuation caused by floors. The floor is concrete with 22 cm thickness. The experiment is done in the second floor. 4 APs are placed on the ground, in the first and third floors, to get the same distance between each AP and a Laptop located in the second floor. RSS measurements are collected at a fixed location, over 5 minutes, with 3 seconds intervals. Figure (8) shows the effect of concrete floors.

TABLE III. FLOORS ATTENUATION

Floor 2,1	15.4 db
Floor 2,3	18.7 db

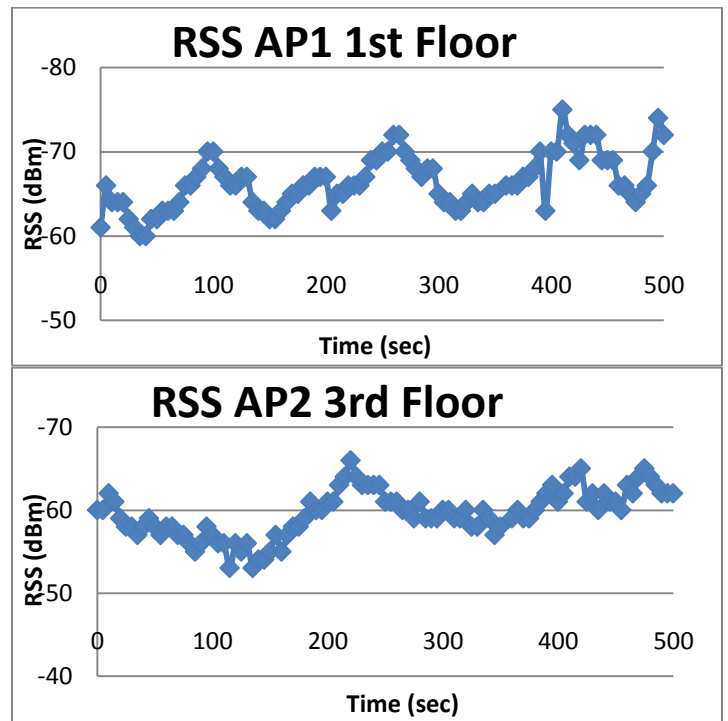


Fig. 8. Effects of concrete floors

VIII. SYSTEM MODEL

Ranging error, which occurs due to obstructions that attenuate the signal, is a major difficulty in indoor localization algorithms. A technique is proposed to overcome this error. The technique works in two modes, which are range-based and range-free modes. The technique first estimates the locality of a user node, which depends on the number of reference nodes in range.

In the second step the locality information is used as input to the second part of the algorithm. The algorithm then collects data about the number and type of obstructions (Table 2 and 3), and finally uses multi wall and floor propagation model in (5) to estimate the location of user node.

The proposed technique requires the storage of the building layout and the attenuation loss of the obstructions (walls and floors). This technique can be implemented in either a centralized manner or in a distributed manner.

In the first step, the algorithm checks whether the user node is in the coverage area of the reference node or not. The check is passed when the distance between the user node and the reference node is less than the maximum radius of the coverage area.

If a user node, in special cases, receives signals from only 2 reference nodes, it can find its locality based on the geometry theory as shown in Figure (9) [8].

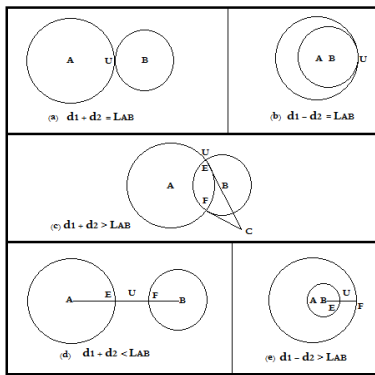


Fig. 9. five cases using two circle to determine the location of the unknown node

When the two circles are externally tangent or internally tangent as shown in Figures 9(a) and (b), the estimated intersection point of the unknown node is taken. If the two circles intersect, there are two intersection points as shown in Figure 9(c). In this case, a revised RSS between the third beacon node and the unknown node as the reference is used to compare  $|d_3 - CE|$  with  $|d_3 - CF|$  and select the smaller one as the estimated unknown node. Here, RSS from the third beacon node is utilized as a reference to identify the proper position between two intersection points.

In in Figures 9(d) and (e), there may be no intersection point because of measurement error. In these cases, we take the middle point of and as the unknown node E F.

### IX. ALGORITHM

- Estimate the locality of the user node from the intersection of the strongest reference nodes signals.  
 For all user node  $l$ ;  
 $Loc\_indicator = 0$ ;  
 For each reference node  $j$   
     If  $(x_j - x_l) + (y_j - y_l) \leq R_j^2$   
      $Loc\_indicator ++1$ ;
- Collect data about the number and type of walls and floors between the estimated user node and each reference node.
- Use the multi wall and floor propagation model to correct the location of user node.

### X. EXPERIMENTAL SETUP

A series of experiments are conducted in the Training center building and the Administration building. Some of the experiments are done in 2-D by keeping the reference and user nodes at the same height, while other experiments are done in 3-D. The first set of experiments is done in the training center building, which has 7 reference nodes, and one user node that captures RSS via Xirrus WiFi inspector software. The location of the user node is estimated using the proposed first part of the algorithm.

The second set of experiments is done in the administration building, which has 12 reference nodes. 6 reference nodes are

on the ground of the first floor, and 6 reference nodes are on the ground of the third floor. The error in the estimated location is calculated as the distance between the actual and estimated locations as expressed in (3).

$$Error = \sqrt{(X_{est} - X)^2 + (Y_{est} - Y)^2} \quad (3)$$

### XI. EXPERIMENT RESULTS

In each experiment, 10 locations are chosen in each building to calculate the error before and after adding the effect of walls and floors. Experimental results show that the position estimation error is less than 2 meters for most locations.

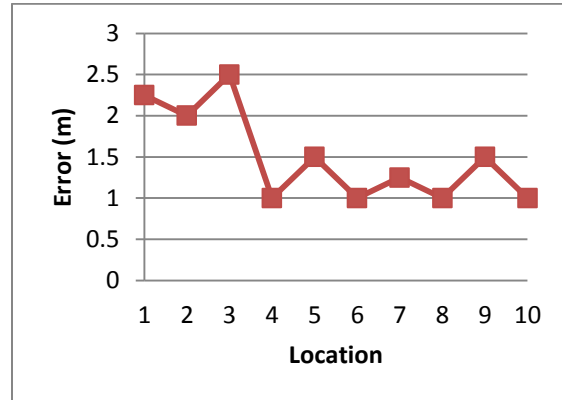


Fig. 10. Location estimation error for various locations in training center building

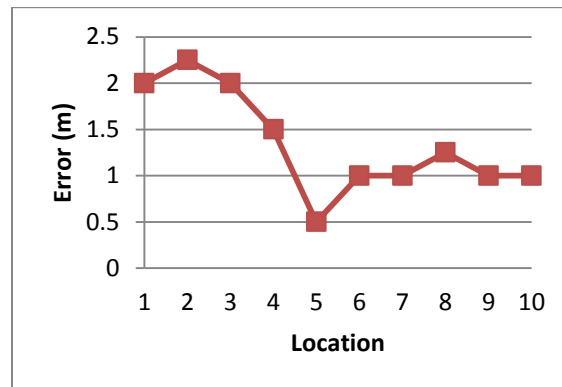


Fig. 11. Location estimation error for various locations in administration building

### XII. CONCLUSION

Currently there is no indoor localization system that can provide an accurate position estimation for all environments. This paper provides an experiment that illustrates the distribution of RSS is presented. A method is then presented to predict the location using RSS. Also, the paper provides a study on the effects of walls and floors on the RSS, and investigates a localization technique that utilizes the finding of the effects of obstructions on the RSS. The investigated technique uses multi walls and floors model in the estimation. The system model showed the major difficulty in indoor localization algorithm, especially the difficulty of location estimation using the storage of building layout, the algorithm has been used to estimate location using three or more

reference nodes, and in special cases, by using two reference nodes. Experimental results show that the position estimation error is less than 2 meters for most locations. It is suitable for indoor environments with multiple floors and multiple walls.

REFERNECE

- [1] Lott, M. and Forkel, I. "A Multi wall and floor for indoor radio propagation", Vehicular Technology Conference, Volume 1, 464-468, 2001.
- [2] Kaemarungsi, K. and P. Krishnamurthy, "Properties of indoor received signal strength for WLAN location fingerprinting", Proceedings of the 1st Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MOBIQUITOUS04), 14-23, 2004.
- [3] Komar, C. and C. Ersoy, "Location tracking and location based service using IEEE 802.11 WLAN infrastructure", European Wireless, 24- 27, 2004.
- [4] Binghao, L. , James, S. ; Andrew, G. D. and Chris, R. "Indoor positioning techniques based on wireless LAN", First IEEE International Conference on Wireless Broadband and Ultra Wideband Communications, Sydney, Australia. 2006.
- [5] Al-Ahmadi, A. S. ; Omer, A. I. ; Kamarudin, M. R. and Rahman, T. A. "multi-floor indoor positioning system using bayesian graphical models", Progress In Electromagnetics Research B, Vol. 25, pp. 241-259, 2010.
- [6] Elnahrawy, E. ; Martin, R.W. ; Ju, P. K. and Madigan, D. "Bayesian indoor positioning systems", Infocom. Citeseer, pp. 1217- 1227, 2005.
- [7] Komar, C. and Ersoy, C. "Location tracking and location based service using IEEE 802.11 WLAN infrastructure", European Wireless, pp. 24-27, 2004.
- [8] Xiao Fan and Yoan Shin, "Indoor Localization for Multi-Wall, Multi-Floor Environments in Wireless Sensor Networks" , AICT 2013.
- [9] Hossain, A. K.; Hien, N. V ; Yunye, J. and Seng, S. W. "Indoor Localization Using Multiple Wireless Technologies", in Mobile Adhoc and Sensor Systems, 2007. MASS 2007. IEEE International Conference, pp. 1-8, 2007.
- [10] Baala, O. ; You, Z. and Caminada, A. "The Impact of AP Placement in WLAN-Based Indoor Positioning System", in Networks, 2009. ICN '09. Eighth International Conference, pp. 12-17, 2009.
- [11] A. Fink, H. Beikirch, M. Voss, and C. Schroeder. RSSI-based Indoor Positioning using Diversity and Inertial Navigation. In IEEE International Conference on Indoor Positioning and Indoor Navigation (IPIN), pages 1-7, 2010.
- [12] Seow, C. and S. Tan, "Localization of omni-directional mobile device in multipath environments", Progress In Electromagnetics Research, Vol. 85, 323{348, 2008.
- [13] Pandey, S. and P. Agrawal, "A survey on localization techniques for wireless networks", Journal Chinese Institute of Engineers, Vol. 29, No. 7, 1125, 2006.

# E-Learning Collaborative System for Practicing Foreign Languages with Native Speakers

Ilya V. Osipov

i2i study, Inc. 2829 Baker St., San  
Francisco, CA 94123, USA

Alex A. Volinsky

Department of Mechanical  
Engineering, University of South  
Florida, 4202 E. Fowler Ave.,  
ENB118, Tampa FL 33620, USA

Anna Y. Prasikova

i2i study SIA  
Krišjāņa Barona Iela, 130 k-10, Rīga,  
Lv-1012, Latvia

**Abstract**—The paper describes a novel social network-based open educational resource for practicing foreign languages with native speakers, based on the predefined teaching materials. This virtual learning platform, called i2istudy, eliminates misunderstanding by providing prepared and predefined scenarios, enabling the participants to understand each other and, as a consequence, to communicate freely. The developed system allows communication through the real time video and audio feed. In addition to establishing the communication link, it tracks the student progress and allows rating the instructor, based on the learner's experience. The system went live in April 2014, and had over six thousand active daily users, with over 40,000 total registered users. Monetization has been added to the system, and time will show how popular the system will become in the future.

**Keywords**—E-learning, learning tools; peer-to-peer network; social network; open educational resources; distance learning

## I. INTRODUCTION

Open educational resources (OER) have recently become quite popular in the area of computer assisted language learning [1, 2, 3]. Currently there are several educational services on the market with a considerable amount of OERs that provide an opportunity to learn foreign languages (livemocha.com [4], www.learn-english-online.org [5], www.duolingo.com [6]). Most of these systems are automated, i.e. don't require live human interaction. These systems can be divided into two categories: autonomous and social. Autonomous methods offer tasks, which are checked or monitored according with the algorithms set up within the system (tests, quizzes, etc., [7]). Social methods allow direct or indirect interaction with real people, including communication, checking assignments, etc. [8]. Such systems have been used in language learning [9, 10, 11, 12]. Tal and Yelenevskaya [13] also attempted to integrate computer-assisted language learning systems into the educational process. One of the currently largest and most popular systems is livemocha.com, which is a social network for the learning foreign languages. The service started in 2007 and currently has over 14 million users. The site uses community learning approach, where students learn through live human interactions using video chat. Currently 38 languages are offered, including four levels for each language. Another system, learn-english-online.org has predefined learning materials, but is complicated to use. There is a chat feature; however, it is not clear how to connect the users through

Skype. Another system, duolingo.com uses predefined materials without live human interaction, which is quite important in foreign language learning and provides better learning outcomes [14, 15].

Currently the most popular distance language learning tool is Skype [16], the real time audio-video communication system, which is not designed for this purpose. Thousands of small companies and individuals offer foreign language learning through Skype. Skype is even being positioned as an instructional resource [17]. Searching for "English via Skype" in Google gives over 43 million results, and similar numbers appear if the search is conducted in Spanish, Russian and other languages. There is a large demand for online foreign language education [18]. However, Skype does not allow finding the person willing to teach/learn foreign languages, it does not provide learning aid materials, and does not track the spent time in the user accounts. Several papers discuss Skype as the language learning tool [16].

The novel method of online collaborative learning of foreign languages takes social learning to a new level, using direct interaction between users, one of which has the competence as a native speaker, and the other one is the recipient of knowledge (the language learner). This virtual learning platform allows to eliminate the problem of misunderstanding each other between all languages, by providing participants with prepared and predetermined scenarios, which enables the participants to understand each other and, as a consequence, to communicate freely. This communication is informal, resulting in better learning outcomes [14, 19]. The service is utilizing task-oriented interactions [21]. It allows overcoming the learner's anxiety, similar to the initial attempts [21], but on a much larger scale, currently supporting four languages.

This system provides an opportunity for the participants to have real time audio and video communication with each other, based on the WebRTC and Adobe Air technologies, followed by interactive prompts, which are controlled by the participants in real time. The prompts consist of images, texts and videos displayed to each participant in their native understandable language. These technological solutions allow using this system with the help of virtually any personal computer with the microphone and the camera connected to the network at the same time without additional specific technical requirements. This methodology can be used to learn



foreign languages from the native speakers in real time, based on the predetermined teaching materials, which simulate real life situations using task-oriented interactions [20]. It is also a great research tool in the area of computer-assisted learning. The paper describes the developed system and ways of viral user attraction with gamification elements.

## II. RESEARCH OBJECTIVES

The i2istudy.com is currently a free multilingual web service for practicing foreign languages online. The main idea of the service is based on the “time banking” principle [22, 23]. For every minute that a person teaches in the mother tongue, s/he is rewarded with a minute that can be used to learn a foreign language. This currently allows using the system free of charge [24].

The name “i2istudy” comes from the idea of the “eye to eye” learning, based on the peer2peer principle [25, 14]. A new model of studying with the native speaker according to a set of interactive courses was created. Every “lesson” is based on the split screen platform (Figure 1). On one half of the screen the native language teacher is shown, and on the other half is a set of audio and video slides (Figure 1). Together with the native speaker, the learner goes through the predefined content. Therefore, there is no need for the professional teacher, and theoretically every person can teach their mother tongue within this format. Furthermore, the i2istudy is a self-regulated system, as it allows every user to pick their own paste, time, level, gender and plenty of other characteristics of their instructor and the lesson, thus everyone can find their “right” teacher.

The i2istudy is an extremely social and informal approach to learning languages [14]. It is not only a mere learning platform, but also a learning community that brings users together and builds relationships [26]. Learning foreign languages is a universal international factor uniting like-minded individuals all over the World. While studying, the system users enter into an intense communication process with a native speaker, not only studying the language, but also the culture, behavior and the manners. They not only discuss the set topics, but can also enter into a more personal communication. At the present time the system supports four languages: English, Spanish, Russian and German, however more languages can be added without significant technical changes of the system [1].

This virtual learning platform provides users with the opportunity to discuss particular subjects in situations where one of the users is a native speaker and an expert in the particular knowledge and the other party is the recipient of this knowledge. Each user is provided accompanying materials in accordance with the level of competence in the subject. Communication in different languages between users is made possible with the help of predefined communication scenarios for the purpose of studying a foreign language.

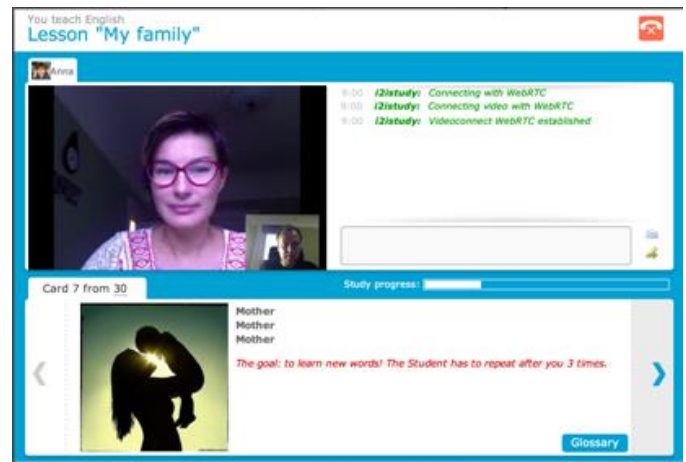


Fig. 1. i2istudy lesson user interface

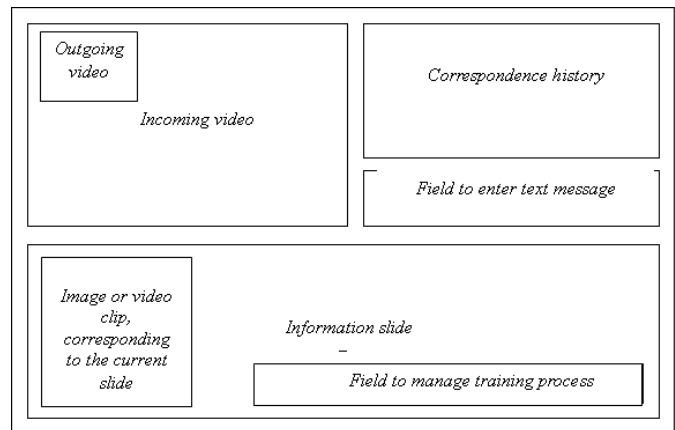


Fig. 2. System user interface layout schematics

This capacity is provided through the combination of three components for the user interaction with each other: one component provides real time audio-visual communication for users, the next component displays text and a scenario of topics in stages used for communication in foreign languages understood by the users, and the third component allows communicating by sending and receiving instant text messages (Figure 2).

The system user interface is shown in Figure 1 and schematically in Figure 2. The system automatically tracks the time, which is reported in user accounts for the purpose of time banking [27]. The user interface, user interaction and the script are presented by means of individual cards (slides), connected by a mutual discussion topic. Each slide consists of a separate text, graphics and video in an interface, which is clear and understandable to each user, in their own native language (Figure 2). Each slide consists of a set of common fields provided in multiple languages, or in each individual language (Figure 3).

The instructor, while being the one responsible to issue comments, dictates to the student what and how to speak, in addition to how to respond to his/her statements. The student is displayed with the minimal information required, since his/her task is to try to understand the native speaker with the amount of information provided. In general, it is assumed that the instructor is available to lead the teaching process, namely, start the training, switch the information slides and decide when to stop the lesson, as well as activate the linguistic tips for the student when necessary (Figure 4). The student also has the ability to activate available linguistic tips. For the purpose of studying foreign languages, the instructor switches the slides and reads the tasks, while the student listens and sees the native speaker, receiving minimal amount of information required for basic understanding. If the student does not understand the teacher, the text can be displayed corresponding to what the teacher said, along with its translation. Depending on the user's role (teacher/student), the slides form a coherent set of readily available information for both parties.

In order to implement more complex forms of interaction, it may be necessary to have more roles. The system also allows for the connection of multiple users with individual roles in the educational process. The logic of assigning roles within the framework of the system is that each user separately determines their own solutions. For example, to organize teaching of several students by a specialist in a particular knowledge, along with the monitoring of the teaching process, it may be necessary to have the roles of a teacher, a student and a supervising controller. The teacher conveys the material to one or more students. However, the decision concerning the successful presentation and mastering of the material, for example, whether to jump to the next course or not, is taken by the controller (Figure 5). The controller oversees the educational process. For example, decisions whether the students can independently take the interactive tests available in the system, along with assessing the competency level of the teacher are taken by the controller. Currently, for the study of foreign languages in the form of a game, the system allows only two roles: the teacher and the student. However, there was a logistic problem of how the students and the teachers would be connected, thus a special module was designed, called the connector module.

*The connector module*

The connector functions are summarized in Figure 6. The main obstacle to the users' interaction in real-time is the search for available users in order to begin the interaction at a specific time. While trying to establish a connection with a specific user, which for some reason does not answer an incoming call, other users are deemed unavailable for the call.

This is a classic problem of a phone call: when you call one number, there is no opportunity to call other recipients, since your line is busy, in addition to the fact that the current recipient of the call is also unavailable to other users.

This problem is solved by implementing the multi-call system, allowing one user to send requests to multiple recipients, increasing the response probability. This solution is not burdensome for users, as it allows monitoring the availability of users at the moment, while making the user inaccessible for additional calls. After the line is freed up, the user is given an option to connect missed calls from others who called while the line was busy.

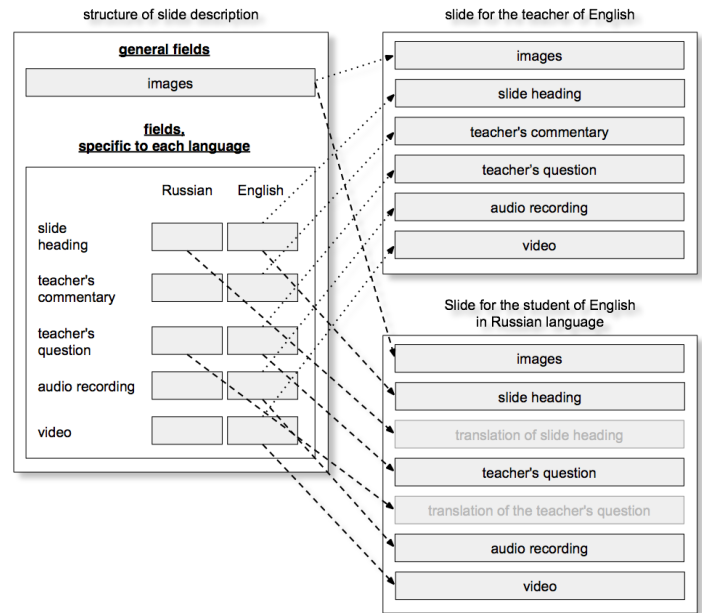


Fig. 3. The slide structure, description and the format, based on the different user roles

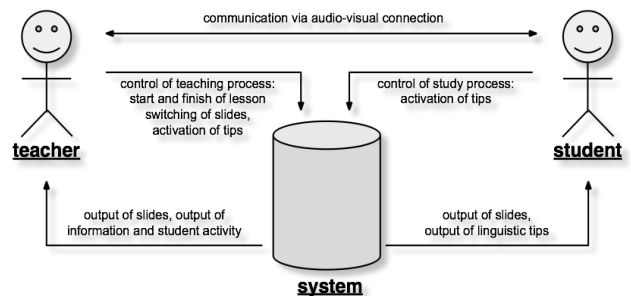


Fig. 4. The teacher - student interaction

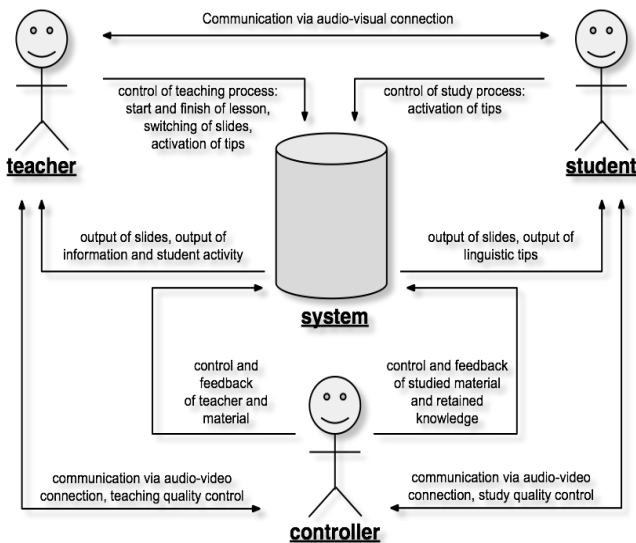


Fig. 5. Teacher-student-controller interaction

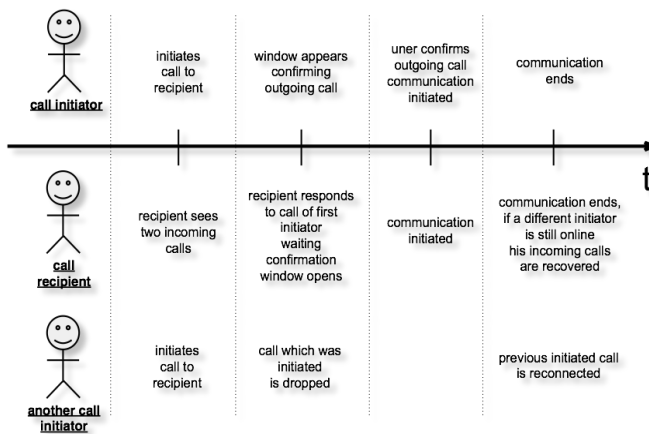


Fig. 6. Connector function diagram

When the user clicks on the notification for the incoming calls, both users receive a notification at the beginning of the call (Figure 7). At the same time both incoming and outgoing calls of other users will go on hold. The user's data and their incoming and outgoing call information become unavailable to all other users. Thus, for the user who answered the call, a notification is sent about the beginning of communication. In other windows and client pages, a notification is displayed that the user is active. The user receiving the call, along with the caller, both receive a notice for the administration of a connection or its cancellation. In order to initiate the connection, the user confirms that s/he is ready; thus communication is established between the clients, generating a response and confirmation.

*Software implementation and initial challenges*

The server portion was written using the PHP programming language combined with the MySQL database. The user client needs an HTML browser with the Java Script installed. Adobe Flash Air or WebRTC are used for the audio-visual real time connection [28].

Initially there were multiple compatibility problems with Adobe Flash Air and WebRTC due to a large number of different browsers and their versions installed by the users. While the initial testing went quite well, there were browsers combinations that could not establish a connection. The number of browser combinations was quite high. The problem was solved by collecting the statistics with further fine tuning of the browser settings and selecting the proper technology supported by the system.

Additionally, there were multiple problems associated with the new users, who didn't understanding how to use the system. The whole concept of the i2istudy.com is quite novel, and initially puzzled the newly registered users. The problem was solved by adding numerous prompts, helping the users navigate through the system. Detailed help was also developed, including frequently asked questions and videos of the demo lessons. The videos of the demo lessons were dubbed in the four languages, currently available in the system (English, Russia, German and Spanish). The users are more likely to follow the steps outlined in the videos, and the amount of questions and misunderstanding was reduced dramatically.

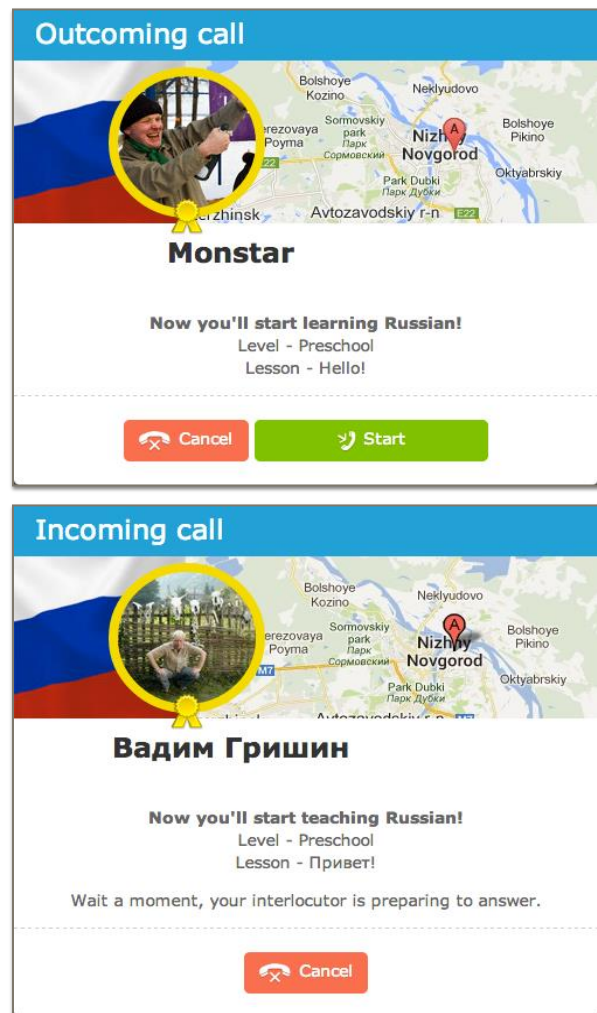


Fig. 7. Answering an incoming call: waiting for the confirmation of the recipient (top) and the initiator of the call confirmation (bottom)

While currently the system requires the use of a personal computer, in the future there could be an option to use it on mobile devices as well, as mobility is the key for the language learning [29].

#### *Gamification technology and virality*

Currently the i2istudy system allows:

Choosing the teacher in the system, based on the offered language, country, gender and age. Sending requests for learning using the Connector module.

Choosing lessons from the library of themes and the language proficiency level.

Tracking the time in minutes spent teaching and learning in the individual user accounts for the purpose of time banking.

Facilitating the teaching/learning process, consisting of the three components: live video, step-by-step learning with the aid of predefined materials in the native language, and the chat window.

Besides these features, other aspects of gamification and virality are being addressed [30]. Gamification is needed to keep the user short term and long term attention, making sure that s/he will come back and continue to use the system in the future [31]. The following attributes are utilized. Numerical value represents the user involvement, along with the leader board. Each user activity is tracked over a month timeframe, allowing the user to compare his/her results with others, based on their respective progress. The rating system for each user allows rating learning/teaching partners after each lesson. The scores are accumulated and are visible to other users, encouraging them to improve their scores. Accessing parts of the lessons is allowed. Passing each set of lessons allows opening new lessons, similar to a computer game, intriguing and entertaining the user. Each lesson progress is tracked with the progress bar, showing how many slides have been studied, and how many slides are still left in the lesson. Information about the newly learned words and phrases is also displayed. Custom decals and badges are awarded to the users, based on their activity in the system. After successfully completing several lessons, the user gets the "Expert" status, which is promoted as an achievement, and is visible to other users. Information about completed lessons can be readily shared in social networks as well. The user gets more involved in the game, achieving higher status with more activity in the system.

Besides keeping the current users involved, the system benefits from attracting the new users, using the following virality mechanisms:

Inviting friends through social networks, like Facebook and others.

Encouraging users to send invitations to his/her friends to attract them to the system. Currently each user gets extra 30 minutes of free lessons for each invited friend, who joins.

Sharing learning/teaching results in social networks if desired.

Detailed description of the system gamification, virality and user retention is a subject of a separate publication [31]. Based in the conducted market analysis, the users demanded system monetization [32], which has been implemented.

### III. CONCLUSIONS

The paper demonstrates the novel approach to learning foreign languages online from the native speakers. The system went live in April 2014, and had over 6,000 active daily users, with over 40,000 registered users. Over 300 new users registered daily. There were over 10 hours of live learning and teaching per day and 12 hours spent with an automated teacher (recorded lessons). These numbers have been growing even without active advertising. However, the project has been closed for three months to add monetization and other system improvements. The developers hope that the i2istudy.com open educational resource will become quite popular with the learners of foreign languages around the World. Future work includes monetization, viral user attraction and retention by adding new lessons and even allowing advanced users create their own lessons, along with supporting more languages. Chinese demo lessons are in the works.

### ACKNOWLEDGEMENTS

The authors would like to thank the i2istudy.com team members for their dedicated efforts: Vadim Grishin, Ilya Poletaev, Andrei Poltanov, Elena Bogdanova, Vildan Garifulin and Franziska Rinke.

### REFERENCES

- [1] Osipov I.V., Prasikova A.Y., Volinsky A.A., (2015). "Participant behavior and content of the online foreign languages learning and teaching platform", *Computers in Human Behavior*, 50, 476-488., DOI: 10.1016/j.chb.2015.04.028.
- [2] Bubnov G., Nikulchev, E., Pluzhnik, E. (2015) "Experience the effective implementation of innovative information technologies in educational institute", *Vysshee obrazovanie v Rossii [Higher Education in Russia]*, 1, 159-161 (In Russian)
- [3] Harasim, L. (2012). *Learning theory and online technologies*, Routledge, New York and London, 191, ISBN 978-0-41-599975-5
- [4] Sevilla-Paóvn, A., Martínez-Sáez, A., Gimeno Sanz, A., Seiz-Ortiz, R. (2012). The role of social and collaborative networks in the development of in-house multimedia language learning materials. *Procedia Social and Behavioral Sciences*, 46, 1826.
- [5] Giles, J. (2012). Learn a language, translate the web. *New Scientist*, 213, 18.
- [6] Rutkin, A. (2014). The next wave of education. *New Scientist*, 222, 27.
- [7] Son, J-B. (2007). Learner experiences in web-based language learning. *Computer Assisted Language Learning*, 20(1), 21-36.
- [8] Yousefi, S. (2014). Comparison of traditional and video mediated learning of English: Tracking a new approach. *Procedia Social and Behavioral Sciences*, 98(6), 1940.
- [9] Donmus, V. (2010). The use of social networks in educational computer-game based foreign language learning. *Procedia Social and Behavioral Sciences*, 9, 1497.
- [10] Kurata, N. (2010). Opportunities for foreign language learning and use within a learner's informal social networks. *Mind, Culture, and Activity*, 17, 382.
- [11] Aydin, S. (2014). Foreign language learners' interactions with their teachers on Facebook. *System*, 42, 155.
- [12] Toetenei, L. (2014). Social networking: a collaborative open educational resource. *Computer Assisted Language Learning*, 27(2), 149.

- [13] Tal, M., Yelenevskaya, M. (2012). Computer-assisted language learning: Challenges in teaching multilingual and multicultural student populations. *Procedia Social and Behavioral Sciences*, 47, 263.
- [14] Lai, K.-W., Khaddage, F., Knezek, G. (2013). Blending student technology experiences in formal and informal learning. *Journal of Computer Assisted Learning*, 29(5), 414-425.
- [15] Blum-Kulka, S., Dvir-Gvirsman, S. (2010). Peer interaction and learning. *International encyclopedia of education (Third Edition)*, 444-449.
- [16] Rao, B., Angelov, B., Nov, O. (2006). Fusion of disruptive technologies: Lessons from the Skype case. *European Management Journal*, 24(2-3), 174.
- [17] Kiziltan, N. (2012). Teaching Turkish through Teletandem. *Procedia Social and Behavioral Sciences*, 46, 3363.
- [18] Kozar, O., Sweller, N. (2014). An exploratory study of demographics, goals and expectations of private online language learners in Russia. *System*, 45, 39.
- [19] Cheong, C., Bruno, V. (2012) Designing a Mobile-app-based Collaborative Learning System. *Journal of Information Technology Education: Innovations in Practice* 11: 97-119.
- [20] Osipov I.V., Volinsky A.A, Nikulchev E., Prasikova A.Y. (2015). "Study of gamification effectiveness in online e-learning systems", *IJACSA Vol. 6, No. 2, 2015, DOI:10.14569/IJACSA.2015.060211*
- [21] Delahunty, J., Verenikina, I., Jones, P. (2013). Socio-emotional connections: identity, belonging and learning in online interactions. A literature review. *Technology, Pedagogy and Education* <http://dx.doi.org/10.1080/1475939X.2013.813405>.
- [22] Válek, L., Jašíková, V. (2013). Time bank and sustainability: The permaculture approach. *Procedia Social and Behavioral Sciences*, 92, 986.
- [23] Seyfang, G., Longhurst, N. (2013). Growing green money? Mapping community currencies for sustainable development. *Ecological Economics*, 86, 65.
- [24] Seufert, E.B. (2014). *Freemium Economics: Leveraging Analytics and User Segmentation to Drive Revenue (The Savvy Manager's Guides)*. Morgan Kaufmann. Waltham MA 02451 USA.
- [25] Hsu, M.-H., Jub, T.L., Yen, C.H., Chang, C.-M. (2007). Knowledge sharing behavior in virtual communities: The relationship between trust, self-efficacy, and outcome expectations. *International Journal of Human-Computer Studies*, 65(2), 153-169.
- [26] Cohen, E.L. (2014). What makes good games go viral? The role of technology use, efficacy, emotion and enjoyment in players' decision to share a prosocial digital game. *Computers in Human Behavior*, 33, 321-329.
- [27] Marks, M. (2012). Time banking service exchange systems: A review of the research and policy and practice implications in support of youth in transition. *Children and Youth Services Review*, 34(7), 1230.
- [28] Karopoulos, G., Mori, P., Martinelli, F. (2013). Usage control in SIP-based multimedia delivery. *Computers & Security*, 39(B), 406.
- [29] Virvou, M., Alepis, E., Troussas, C. (2011). MMALL Multilingual mobile-assisted language learning. *BMSD 2011 - First International Symposium on Business Modeling and Software Design*.
- [30] Fields, T., Cotton, B. (2011). *Social Game Design. Monetization Methods and Mechanics*. Morgan Kaufmann. Waltham MA 02451 USA.
- [31] Osipov I.V., Volinsky A.A., Grishin V.V. (2015) Gamification, virality and retention in educational online platform. Measurable indicators and market entry strategy. *International Journal of Advanced Computer Science and Applications*, 6(4), DOI: 10.14569/IJACSA.2015.060402
- [32] Osipov I.V., Volinsky A.A, Nikulchev E., Plokhov D. (2015) "Study of Monetization as a Way of Motivating Freemium Service Users", *Contemporary Engineering Sciences* 09/2015; 8(20):911-918., DOI:10.12988/ces.2015.57212



# Color Image Segmentation via Improved K-Means Algorithm

Ajay Kumar

Department of Computer Science & Engineering  
Jaypee University of Engineering & Technology-473226,  
Guna (MP), India

Shishir Kumar

Department of Computer Science & Engineering  
Jaypee University of Engineering & Technology-473226  
Guna (MP), India

**Abstract**—Data clustering techniques are often used to segment the real world images. Unsupervised image segmentation algorithms that are based on the clustering suffer from random initialization. There is a need for efficient and effective image segmentation algorithm, which can be used in the computer vision, object recognition, image recognition, or compression. To address these problems, the authors present a density-based initialization scheme to segment the color images. In the kernel density based clustering technique, the data sample is mapped to a high-dimensional space for the effective data classification. The Gaussian kernel is used for the density estimation and for the mapping of sample image into a high-dimensional color space. The proposed initialization scheme for the k-means clustering algorithm can homogeneously segment an image into the regions of interest with the capability of avoiding the dead centre and the trapped centre by local minima phenomena. The performance of the experimental result indicates that the proposed approach is more effective, compared to the other existing clustering-based image segmentation algorithms. In the proposed approach, the Berkeley image database has been used for the comparison analysis with the recent clustering-based image segmentation algorithms like k-means++, k-medoids and k-mode.

**Keywords**—k-means; k-means++; k-medoids; k-mode; kernel density component

## I. INTRODUCTION

Unsupervised color image segmentation is an important image processing technique that has a wide application in computer vision applications, pattern recognition, image retrieval [1], image editing [2], and medical image analysis [3]. The objective of image segmentation is to partition an image into the homogeneous region on the basis of an application [4].

Image segmentation algorithms that are based on the clustering can be subdivided into the hierarchical and partitioned techniques. The hierarchical clustering is a bottom-up approach, where a nested cluster structure is obtained by merging the nearby data points. The partitioning clustering is an iterative partitioning process which uses k seed value as an input from the user and each object of the data set must be assigned to precisely one cluster [5]. Due to simplicity and ease of implementation, the k-means clustering [6] and partitioning around medoids [7] are the popular choices for the performing image segmentation.

The k-means algorithm uses the feature of an image to find the k number of groups. The k-means algorithm aims to

minimize an objective function [8], in order to find the groups. For the dataset  $X=\{X_1, X_2, \dots, X_n\}$  with n observations, the purpose of k-means clustering is to find the k groups in X as  $C=\{C_1, C_2, \dots, C_k\}$  such that the objective function  $f_{km}$  is minimized as shown in “(1)”:

$$f_{km} = \sum_{i=1}^n \sum_{k=1}^k z_{in} \|X_i - \mu_k\|^2 \quad (1)$$

Here,  $z_{in}$  is a variable defined in “(2)”.

$$z_{in} = \begin{cases} 1 & \text{if } X_i \in C_k \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Here,  $C_k$  represents the  $k^{\text{th}}$  cluster and  $\mu_k$  represents the mean vector of the observation  $C_k$ .

The main advantage of k-means is that it always finds local optima for any given initial centroid locations. Despite being used in a wide array of application, the k-means algorithm is not exempt from limitations. From a practical point of view, the seed value of the algorithm is vital since each seed can produce different local optima leading to the varying partitions. The quality and efficiency of the algorithm can vary far away from the global optimum, even under repeated random initialization. Therefore, a good initialization is critical for finding the globally optimal partitions. Several methods have been documented in the literature on improving the initialization procedure that changes the performance, both in terms of quality and convergence properties.

In this paper, a density-based color image segmentation technique is used to improve the results of classical partition-based image segmentation methods, like the k-means clustering algorithm. The k-means algorithm is enhanced, by providing a reduced-set representation of kernelized center as an initial seed value. In the kernel density based clustering technique, the data sample is mapped to a high-dimensional space for effective data classification [9]. One of the popular choices is the Gaussian kernel, which is used in the proposed scheme for mapping sample image into a high-dimensional color space. Moreover, a reduced-set kernelized center can be employed for reducing the computational complexity of various algorithms.

The experimental results were compared by using the four types of evaluation measures: Probabilistic Rand Index (NPR), Global Consistency Error (GCE), and Variation of Information (VOI) on the Berkeley image database. The performance of the experimental result indicates that the proposed approach is more effective as compared to other existing clustering-based



image segmentation algorithms such as the k-means, k-means++, k-medoids and k-mode.

The rest of the paper is framed as follows: Section 2 describes the related works. Section 3 provides the proposed approach for Image segmentation via a density-based initialization of k-means algorithm and validation measures. Result and comparative analysis are discussed in Section 4. Finally, the conclusion has been presented in section 5.

## II. RELATED WORK

The k-means class of algorithms suffers from the random selection of initial cluster centers. The arbitrary choice of initial cluster centers leads to the non-repeatable clustering results that may be difficult to comprehend. The results of partition-based image segmentation algorithms are better when the initial partitions are close to the final solution. A short review of the existing work is included in this section for clustering-based image segmentation and computing an initial seed value for the k-means algorithm which is used for the color image segmentation.

T. Pavlidis in 1982 shows the image segmentation process from a wide perspective. It summarizes the use of different methods such as clustering, edge-based segmentation, graph-based approaches, region growing, probabilistic or Bayesian approaches and other approaches for the image segmentation [10].

Chan, *et al.* in 2001 introduced a region-based method known as Chan–Vese or CV model [11]. This method formulates the image segmentation problem as a k-means clustering model. As pointed out in [12], a global method, the CV model cannot solve the intensity irregularity problem in a better way. Wang *et al.* in 2010 pointed out that the local binary fitting (LBF) method is sensitive to the initialization. To prevent this, they introduced the local order method [13]. Besides improving global methods into local versions, many researchers focus on the convexity of the segmentation models.

Arthur and Vassilvitskii in 2007 introduced the k-means++ algorithm to find the initial centers with probability proportional to the distance to the nearest center [14]. Maitra, in 2009, used the local modes present in the data set to initialize the k-means algorithm which is used for the segmentation [15].

In the k-medoids [7] methods, a cluster is represented by one of its points. N-medoids are selected from the given data and clusters are defined as the subset of points close to the respective medoids. Two early versions of the k-medoids methods, the partitioning around medoids (PAM) algorithm and the clustering large applications (CLARA) are a popular choice for the image segmentation. PAM is an iterative optimization that combines the relocation of the points between the perspective clusters with re-nominating the points as potential medoids.

Zhiding *et al.* in 2010 documented an adaptive unsupervised method for the color image segmentation. The algorithm clusters the pixel in 3D, RGB color space by using the ant colony-fuzzy c-means hybrid algorithm (AFHA), which

uses an ant system for intelligent initialization of the cluster centroids [16].

Khan *et al.*, in 2013, introduced a novel initialization scheme to determine the number of clusters and obtain the initial cluster centers for the fuzzy C-means algorithm to segment any kind of color images. The hierarchical approach has been used to integrate the splitting and merging techniques in order to obtain an initialization for FCM [17].

S. Khan *et al.*, in 2013, presented a solution for the randomized initialization of the k-mode algorithm. A prominent attribute selection method has been used to find an initial cluster center. It performs multiple clustering of the data based on the attribute values to obtain a deterministic mode. These modes are used for initialization [18].

The k-modes [19] algorithm allows the user to work with a kernel density estimate of bandwidth “ $\sigma$ ” but produces exactly k clusters. It finds the centroids that are valid patterns and lie in the high-density area. The k-modes algorithm uses local bandwidth at each point rather than a global one.

A good initialization scheme will improve the results of clustering. Thus, following in the same direction, a new initialization technique for color image segmentation has been proposed. The k-mean, k-means++, k-medoids, and k-modes are cumulatively used to demonstrate the effectiveness of the proposed approach.

## III. PROPOSED METHOD

The pixel of a color is represented by the three values corresponding to the R (red), G (green), and B (blue). By using either linear or nonlinear transform on the RGB scale, one can find the color models such as intensity, saturation, and hue.

Each color space has its own characteristic. In the color-based clustering technique, it is desirable that the selected color features are defined in a uniform color space [20]. In order to get the uniform color space, kernel density estimation has been used for estimating the probability density function of a continuous random variable [21]. In the proposed method, a Gaussian kernel based initialization scheme for color image segmentation is used. Unlike the standard k-means, the proposed algorithm uses a density estimate to select an initial cluster center from the color space.

In the following section, the algorithm to select initial seed value for the k-means algorithm that has a high impact on the color image segmentations is presented. These initial points are selected from the denser region of the data sets.

The algorithm starts by choosing the attribute value in a  $n \times m$  data matrix, having maximum variance. A Gaussian kernel is placed over each data point of the selected attribute for the estimation of density. Further, the first seed point is selected, where the density is maximum. The next probable point is selected, that has a density equivalent to the initially selected point. In this way, the k points are obtained, having a similar density with respect to each other.

The kernel density technique is used to estimate the probability density function of a continuous random variable.

Let  $\{X_1, X_2, \dots, X_n\}$  be a sample from a variable P, then the kernel density estimate is a "sum of n kernel functions". In this paper, the popular Gaussian kernels have been used.

Each Gaussian kernel function is centered on a sample data point with the variance h, which is defined as the bandwidth and it is used to control the level of smoothing. The density of the data points depends on the width of Gaussian kernel, so a proper value of h is obtained from the Silverman approximation rule for which,  $h = 1.06 \times \delta |P|^{-1/5}$ , where  $\delta$  is the standard deviation of the sample points P [22].

In one-dimensional case, the density estimator is defined as follows:

$$P(x) = \frac{1}{|P| \sqrt{2\pi h}} \sum_{p \in P} e^{-\frac{[x-p]^2}{2h^2}} \quad (3)$$

For the d-dimensional case, the kernel function is the product of d Gaussian functions, each with its own bandwidth  $h_j$  henceforth; the density estimator is defined as follows:

$$P(x) = \frac{1}{|P| [2\pi]^{d/2} \prod_{j=1}^d h_j} \sum_{p \in P} \prod_{j=1}^d e^{-\frac{(x_{D_j} - p_{D_j})^2}{2h_j^2}} \quad (4)$$

Where a d-dimensional point p is denoted by  $\{P.D_1, P.D_2, \dots, P.D_d\}$ .

The proposed algorithm is briefly described in the following steps:

- Step1.** The data set is first normalized.
- Step2.** The attribute which has the maximum variance is selected.
- Step3.** The density estimate for the selected attribute is computed by using the Gaussian window. For multidimensional data, the kernel function is computed as the product of d Gaussian function, each with its own bandwidth  $h_j$  and the density estimator is given in "(4)".
- Step4.** The first point is selected where the density is maximum. The next (k-1) points are selected from the other denser regions of the data set, so as the density of the points are similar or equivalent with respect to each other.
- Step5.** The indices of selected k data points are used for the initialization purpose.
- Step6.** The k-means algorithm is executed with the help of initial seed value, computed in the previous step.

The motive behind the above method is to find the initial points from the denser area of the given dataset. In this way, the selected data points represent the common characteristics of the entire dataset and are used for the initialization.

The modified k-means algorithm now used to segment any color image. A color image is passed to the modified algorithm and a suitable value of the number of a segment is also set to get the desired segmented image. In proposed algorithm, the images from the Berkeley Image Database [26] are used to test

the validity of the proposed method. Nine images are used though it possible to test it over all the images.

#### A. Validation Measures

The Normalized Probabilistic Rand (NPR) index, Variation of Information (VOI), Global Consistency Error (GCE), and peak signal to noise ratio (PSNR) is used as the validity measure to check the quality of the segmented image. It is important to evaluate the quality of a segmented image obtained by various clustering algorithm because the results of various clustering algorithms gives different results. The NPR index [23] is the generalized version of rand index, which is used to measure the quality of clustering results. The NPR uses the hand-labeled set of ground-truth segmentation to perform a comparison between two image segmentation algorithms. The value of NPR is in the range of -1 to 1, where the high value indicates better segmentation.

The GCE measures the consistency level between the outputs of two segmentation algorithm applied to a given image. It also shows whether there is a refinement relation between the two segments or a possible overlap of pixels. The range of the GCE is between 0 and 1. A value close to zero represents better segmentation.

The Variation of Information metric defines the information gain or information loss between the two segments. It also measures the degree of randomness in the given segment. The range of VOI is  $[0, \infty]$ , a smaller value indicates better results [24]. The PSNR value represents the region homogeneity between image and its segmented image. The higher value indicates the better segmentation results.

## IV. RESULTS AND DISCUSSION

For each algorithm, its correctness is measured by the NPR index, GCE, VOI, and PSNR as well as its stability with respect to changes in the parameter settings and with respect to the different images. An algorithm which produces correct segmentation results with a wide array of parameters on any image, as well as accurate segmentation results on multiple images with the same parameters, will be useful for the pre-processing step in a larger system.

The results are based on the Berkeley image segmentation database [25], which contains 300 natural images along with several ground truth hand segmentations for each image [26]. In contrast to the results presented in this database, the proposed algorithm uses the same image feature (position and color) for segmentation, thereby making their output directly comparable.

The proposed method can be applied directly to the  $321 \times 481$  images taken from the Berkeley database. Due to a few homogeneous groups in an image, it is efficient to reduce the image size for the computation of centroids. Also the MATLAB implementations of k-mode and k-medoids clustering encounter memory errors for the high value of image resolution (running 32-bit MATLAB on a machine with 2GB RAM). Hence, the images are downsized to  $64 \times 64$ .

Nine images from Berkeley image segmentation database are used. The images are segmented (Figure "a" to Figure "i") for  $k=2$  and  $k=3$  value, see fig. 1.

	Original Image	Segmented image	Test Image	k-means	k-means++	k-medoids	k-mode	Proposed
a								
b								
c								
d								
e								
f								
g								
h								
i								

Fig. 1. Segmentation results (Each row from left to right: original image, human segment, test image, k-means based segmentation, k-means++ based segmentation, k-medoids based segmentation, k-mode based segmentation, proposed algorithm)



Four validation tests have been performed for comparative analysis on each image (NPR = Probabilistic Rand Index, VOI = Variation of Information GCE = Global Consistency Error and PSNR).

The Value  $k=3$  represents the number of color segments present in an image. The value  $k=3$  is used because the database itself segmented it into three regions. In order to compare the results of the proposed method with the human segmented image, the value of  $k$  as three is used. That is why the value of  $k=3$  is used for an image a, d, e, f, g, i and  $k=2$  for the image b, c, and h.

In Figure 1, the first column is the true images, the second column is for the human segmented images, the third column contains the test images, and all other columns are the output of the various algorithm used in the paper.

The result of comparative analysis of recent clustering based image segmentation is presented in the Tables 1-9. A histogram comparison of the validation test is also included for better visualization; see Figures 2-10.

After analysing the results, it is also observed that the proposed approach (kernel) performed better compared to the recent clustering based color image segmentation. The NPR results indicate that the proposed algorithm is better in 8 out of nine images used for segmentation and it is equivalent to an image (h), which can be seen in Tables 1-9 and Figures 2-10. The PSNR results show that the algorithm is better in 8 out of nine cases with respect to other algorithm used for comparison. Only for the image (f), the proposed algorithm gives poorer results. The GCE and VOI validation test shows that the proposed method is better in 7 out of 9 cases and equivalent in two cases, see Tables 1-9 and Figures 2-10. It can also be observed that the proposed algorithm gives significantly better results in terms of PSNR and NPR but for other two validation measure results are not significant. The NPR and PSNR value indicate that the proposed method is useful for color image segmentation.

TABLE I. VALIDATION TEST RESULTS ON IMAGE (A)

	k-means	k-means++	k-medoids	k-mode	Proposed
NPR	0.291	0.290	0.297	0.291	<b>0.300</b>
VOI	0.229	0.230	0.229	0.230	0.220
GCE	5.256	5.303	5.256	5.416	5.244
PSNR	3.784	4.009	3.784	4.773	5.856

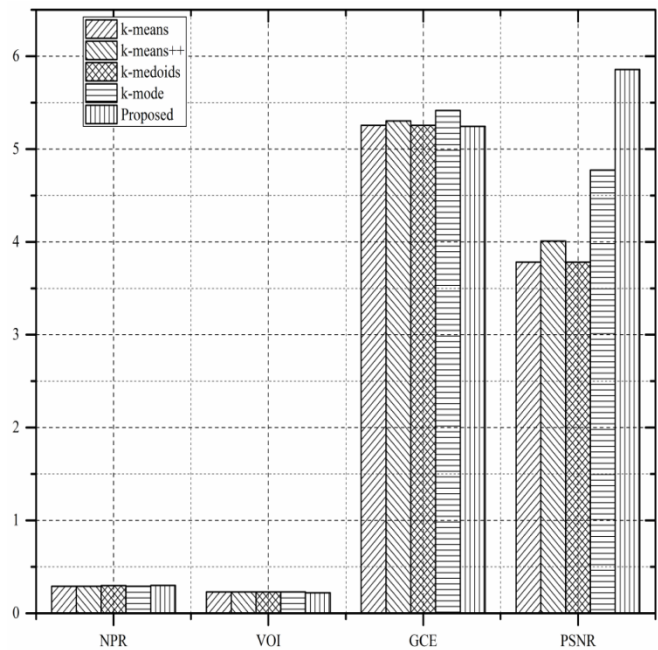


Fig. 2. Comparison of validation results on image (a) for  $k=3$

TABLE II. VALIDATION TEST RESULTS ON IMAGE (B)

	k-means	k-means++	k-medoids	k-mode	Proposed
NPR	0.375	0.374	0.375	0.334	<b>0.385</b>
VOI	0.259	0.258	0.259	0.258	0.258
GCE	3.851	3.873	3.850	4.368	3.850
PSNR	2.748	3.141	2.748	2.908	3.747

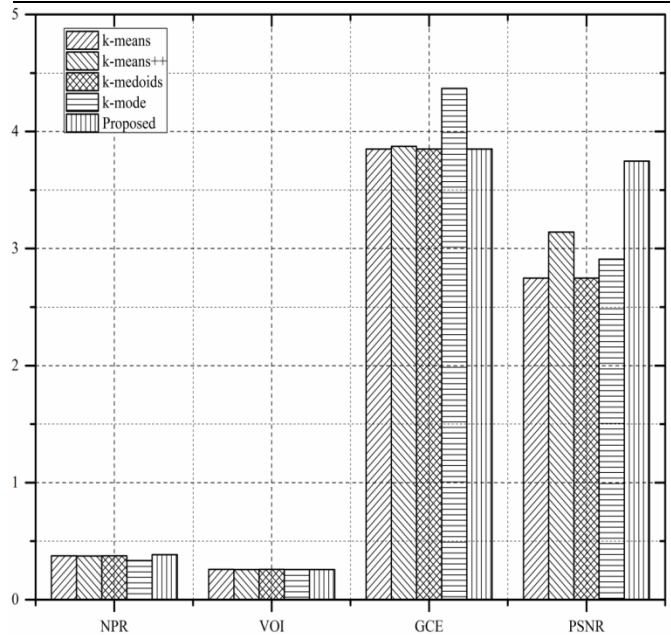


Fig. 3. Comparison of validation results on image on image (b) for  $k=2$

TABLE III. VALIDATION TEST RESULTS ON IMAGE (C)

	k-means	k-means++	k-medoids	k-mode	Proposed
NPR	0.317	0.317	0.316	0.318	<b>0.319</b>
VOI	0.246	0.245	0.245	0.245	0.245
GCE	4.427	4.414	4.427	4.410	4.404
PSNR	2.909	2.873	2.908	2.863	2.911

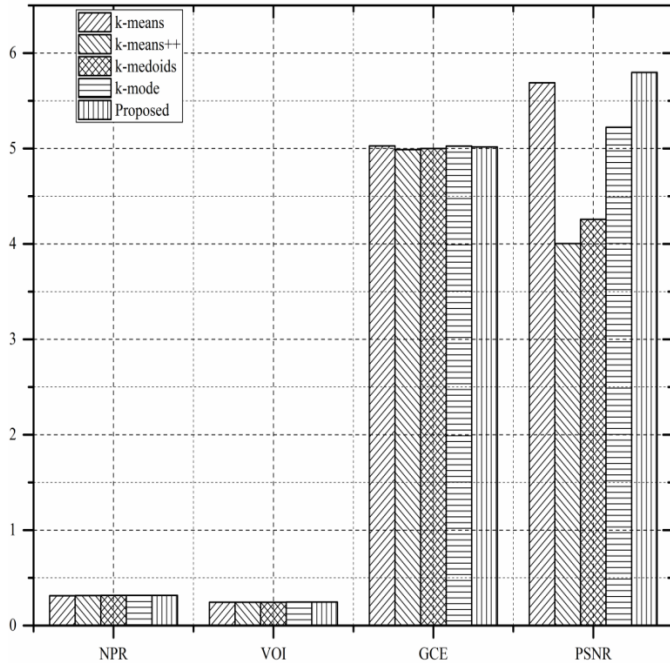


Fig. 4. Comparison of validation results on image on image (c) for k=2

TABLE IV. VALIDATION TEST RESULTS ON IMAGE (D)

	k-means	k-means++	k-medoids	k-mode	Proposed
NPR	0.313	0.315	0.316	0.316	<b>0.317</b>
VOI	0.245	0.245	0.245	0.246	0.246
GCE	5.028	4.989	5.000	5.025	5.016
PSNR	5.690	4.003	4.257	5.222	5.797

TABLE V. VALIDATION TEST RESULTS ON IMAGE (E)

	k-means	k-means++	k-medoids	k-mode	Proposed
NPR	0.323	0.323	0.323	0.318	<b>0.325</b>
VOI	0.259	0.259	0.259	0.258	0.259
GCE	4.865	4.856	4.875	5.527	4.823
PSNR	5.214	3.707	4.642	3.897	5.373

TABLE VI. VALIDATION TEST RESULTS ON IMAGE (F)

	k-means	k-means++	k-medoids	k-mode	Proposed
NPR	0.372	0.375	0.374	0.373	<b>0.380</b>
VOI	0.232	0.232	0.233	0.232	0.233
GCE	3.398	3.371	3.373	3.408	3.341
PSNR	7.077	2.977	7.422	6.978	4.263

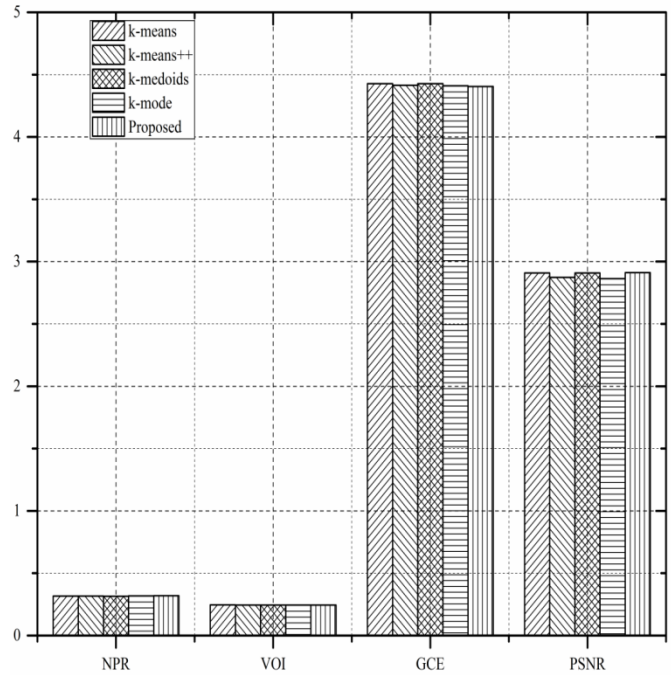


Fig. 5. Comparison of validation results on image on image (d) for k=3

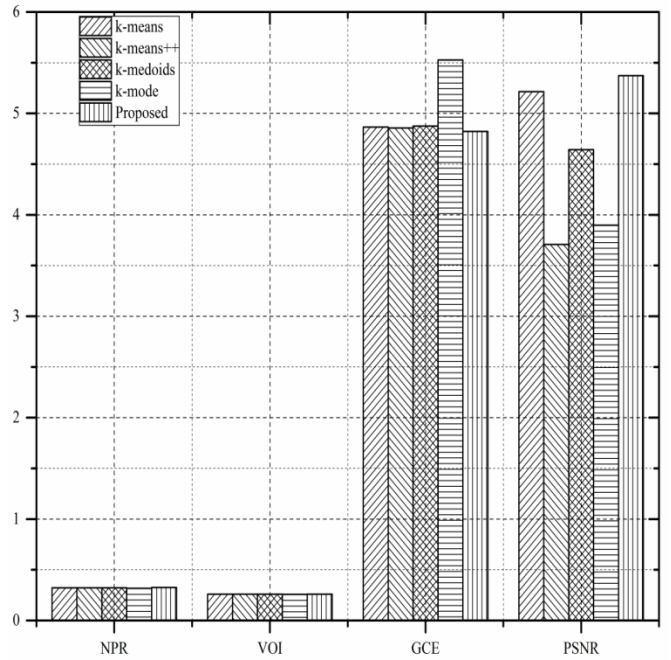


Fig. 6. Comparison of validation results on image on image (e) for k=3

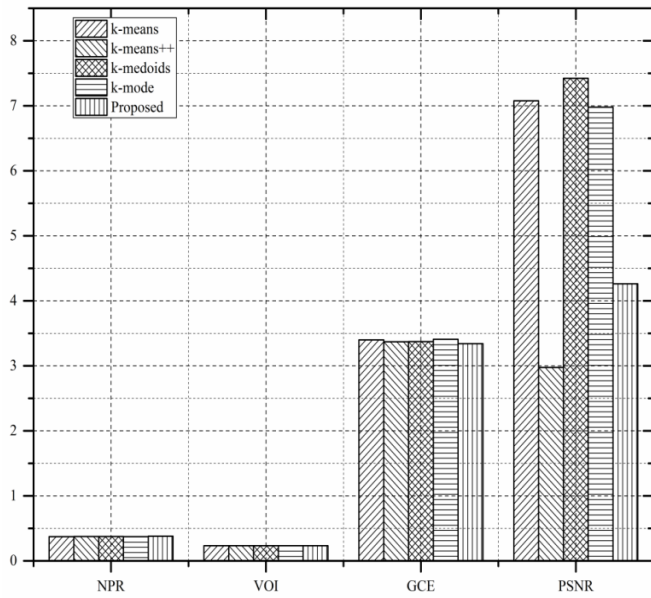


Fig. 7. Comparison of validation results on image on image (f) for k=3

TABLE VII. VALIDATION TEST RESULTS ON IMAGE (G)

	k-means	k-means++	k-medoids	k-mode	Proposed
NPR	0.285	0.262	0.283	0.284	<b>0.286</b>
VOI	0.192	0.191	0.191	0.192	0.191
GCE	4.304	5.025	4.286	4.383	4.277
PSNR	5.198	4.433	3.771	4.467	5.368

TABLE VIII. VALIDATION TEST RESULTS ON IMAGE (H)

	k-means	k-means++	k-medoids	k-mode	Proposed
NPR	0.504	0.271	0.270	0.500	<b>0.504</b>
VOI	0.165	0.166	0.166	0.165	0.164
GCE	2.125	3.940	3.940	2.1655	2.121
PSNR	2.478	2.817	2.863	3.232	3.478

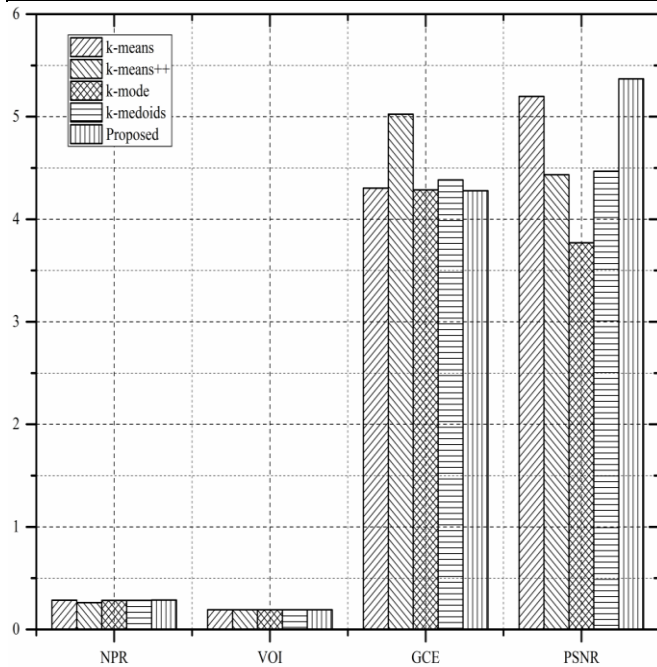


Fig. 8. Comparison of validation results on image on image (g) for k=3

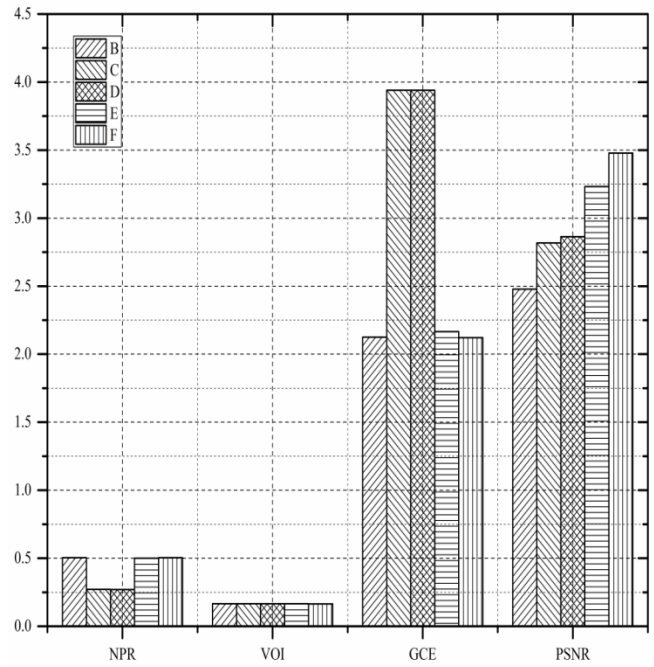


Fig. 9. Comparison of validation results on image on image (h) for k=2

TABLE IX. VALIDATION TEST RESULTS ON IMAGE (I)

	k-means	k-means++	k-medoids	k-mode	Proposed
NPR	0.393	0.393	0.393	0.383	<b>0.395</b>
VOI	0.326	0.326	0.326	0.327	0.326
GCE	4.807	4.796	4.852	4.972	4.852
PSNR	3.872	3.617	5.860	5.310	5.880

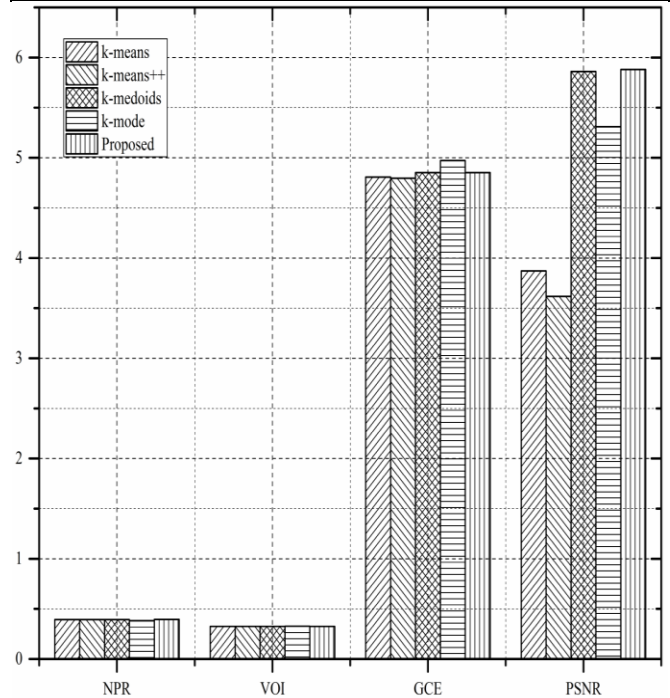


Fig. 10. Comparison of validation results on image on image (i) for k=3

Numerical experiments on the images from the Berkeley database shows that the proposed method is able to perform competitively against the popular clustering-based image segmentation algorithms and often give a close solution with

respect to human segmented images. Therefore, the proposed technique can be used for effective image segmentation.

## V. CONCLUSIONS

A density-based algorithm for initializing the k-means algorithm has been proposed and used in the color image segmentation. Four popular clustering-based image segmentation techniques are used for comparison of the results. The list does not include every possible strategy proposed in the literature. Indeed, it is not practical to compare every method available. However, the work provides a starting point in refining and evaluating new strategies for the k-means algorithm used in the image segmentation.

The Image is first segmented with the k-means algorithm and then with k-means++, k-medoids, k-mode and at last with the proposed method. The results of the segmentation are compared with the help of four validation measures. On the performance basis, the proposed method is better than the k-means and other partitioned based segmentation techniques. The only difference between these techniques is the way of getting initial seed pixel: in the k-means random selection is used while in k-means++ pixels are generated by the weighted probability distribution of the spectrum while the proposed method uses the density of the pixel.

## REFERENCES

- [1] Santhosh P. Mathew, Valentina E. Balas, Zachariah K. P, Philip Samuel, "A Content-based Image Retrieval System Based on Polar Raster Edge Sampling Signature," Acta Polytechnica Hungarica, Journal of Applied Sciences, Vol. 11, No. 3, 2014, pp. 25-36.
- [2] C. Jaisin, S. Pathaveerat and A. Terdwongworakul, "Determining the Size and Location of Longans in Bunches by Image Processing Technique," Maejo International Journal of Science and Technology, 2013, pp. 444-455.
- [3] Hind R.M Shaaban, Farah Abbas Obaid and Ali Abdulkarem Habib, "Performance Evaluation of K-Mean and Fuzzy C-Mean Image Segmentation Based Clustering Classifier" International Journal of Advanced Computer Science and Applications (IJACSA), 6(12), 2015, pp. 176-183.
- [4] CW Bong and M Rajeswari, "Muti-objective Nature-inspired Clustering and Classification Techniques for Image Segmentation," Applied Soft Computing, 2011, pp. 3271-3282.
- [5] Krisztián Kósi, "Method of Data Center Classifications," Acta Polytechnica Hungarica Vol. 9, No. 5, 2012, pp. 127-137.
- [6] J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observation," Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281-297.
- [7] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York, 1990.
- [8] Paul S. Bradley and Usama M. Fayyad, "Refining Initial Points for K-Means Clustering," In Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998, pp. 91-99.
- [9] Liang Liao, Yanning Zhang, "MRI Image Segmentation Based on Fast Kernel Clustering Analysis," Frontiers of Electrical and Electronic Engineering in China, 2011, Volume 6, Number 2, 2011, pp. 363-373.
- [10] T. Pavlidis, Algorithm for Graphics and Image Processing, Computer Science Press, Berlin, 1982.
- [11] T. F. Chan and L. A. Vese, "Active Contours without Edges," IEEE Transaction Image Processing, 2001, pp. 266-277.
- [12] Chunming Li, Chiu-Yen Kao, J.C. Gore and Z. Ding, "Minimization of Region-Scalable Fitting Energy for Image Segmentation," IEEE Trans. Image Process., 2008, pp. 1940-1949.
- [13] L. Wang, Z. Yu, and C. Pan, "Medical Image Segmentation Based on Novel Local Order Energy," Proceedings of the 10th Asian Conference on Computer Vision, Queenstown, New Zealand, 2010, pp. 148-159.
- [14] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007, pp. 1027-1035.
- [15] R. Maitra, "Initializing Partition-Optimization Algorithm," IEEE/ACM Transaction on Computational Biology and Bioinformatics, 2009, pp. 144-157.
- [16] Zhiding Yu, Oscar C. Au, Ruobing Zou, Weiyu Yu and J. Tian, "An Adaptive Unsupervised Approach Toward Pixel Clustering and Color Image Segmentation," Pattern Recognition, 2010, pp. 1889-1906.
- [17] K. S. Tan, W. H. Lim, N. A. M. Isa, "Novel Initialization Scheme for Fuzzy C-Means Algorithm on Color Image Segmentation," Applied Soft Computing, 2013, pp. 1832-1852.
- [18] S. S. Khan and A. Ahmad, "Cluster Center Initialization Algorithm for K-Mode Clustering, Expert Systems with Applications," 2013, pp. 7444-7456, 2013.
- [19] L. Bai, J. Liang, "The K-Modes Type Clustering Plus Between-Cluster Information for Categorical Data," Neurocomputing, 2014, pp.111-121.
- [20] H. D. Cheng, X. H. Jiang, Y. Sun and J. Wang, "Color Image Segmentation: Advances and Prospects," Pattern recognition, 2001, pp. 2259-2281.
- [21] D. Scott, Multivariate Density Estimation, Theory, Practice and Visualization, Wiley, 1992.
- [22] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, 1986.
- [23] R. Unnikrishnan, C. Pantofaru and M. Hebert, "A Measure for Objective Evaluation of Image Segmentation Algorithms," CVPR workshop on Empirical Evaluation Methods in computer vision, 2005, pp. 394-400.
- [24] M. Meilă, "Comparing Clustering: An Axiomatic view," Proceedings of the 22nd International Conference on Machine learning. ICML 05, New York, 2005, pp. 577-584.
- [25] D. Martin, C. Fowlkes, D. Tal and J. Malik, "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," Proceedings of 8th Intl. Conf. Computer Vision, 2001, pp. 416-423.
- [26] D. Martin and C. Fowlkes, Retrieved from <http://www.cs.berkeley.edu/projects/vision/grouping/segbench>, Accessed: October 2015.



# Detection and Feature Extraction of Collective Activity in Human-Computer Interaction

Ioannis Karydis  
Dept. of Informatics  
Ionian University  
Corfu, Greece

Phivos Mylonas  
Dept. of Informatics  
Ionian University  
Corfu, Greece

Markos Avlonitis  
Dept. of Informatics  
Ionian University  
Corfu, Greece

Spyros Sioutas  
Dept. of Informatics  
Ionian University  
Corfu, Greece

**Abstract**—Time-based online media, such as video, has been growing in importance. Still, there is limited research on information retrieval of time-coded media content. This work elaborates on the idea of extracting feature characteristics from time-based online content by means of users' interactions analysis instead of analyzing the content itself. Accordingly, a time series of users' activity in online media is constructed and shown to exhibit rich temporal dynamics. Additionally it is demonstrated that is also possible to detect characteristic patterns in collective activity while accessing time-based media. Pattern detection of collective activity, as well as feature extraction of the corresponding pattern, is achieved by means of a time series clustering approach. This is demonstrated with the proposed approach featuring information-rich videos. It is shown that the proposed probabilistic algorithm effectively detects distinct shapes of the users' time series, predicting correctly popularity dynamics, as well as their scale characteristics.

**Keywords**—Users' activity; aggregation modelling; collective intelligence; time-based media; pattern detection

## I. INTRODUCTION

Over the last decade the World Wide Web became the most popular medium for watching video content [1]. Online multimedia content gains wide acceptance and popularity, due to recent advances in related technology and significant related hardware cost reductions. User-generated content, ranging from personal and how-to videos captured by individuals to free online video lectures and documentaries produced by organizations and academic institutions, is growing at a fast pace online. Broadband Internet connections and increasing interest in online applications and games aided the unprecedented growth of rich multimedia content. Now more than ever more effective methods for indexing, searching, categorizing and organizing this information are required [2].

As a deterministic following, corresponding research work on video retrieval must adopt itself to new techniques and methodologies that leave behind traditional video analysis that is usually based on automatic spatiotemporal analysis of sequences and open up ways to new interpretations. One of the new trends towards this direction is the consideration and

exploitation of the interactive behaviour of users as an integral part of the actual video retrieval process. In this research, this approach is taken a step further and social video consumption activity is considered as a user activity signal within the temporally linear video playback. This approach relies on the capture and analysis of implicit user interactions in order to extract useful information about the video. Previous research [3] has suggested that implicit interactions between users viewing a video and the video-player benefit video summarization. To further explore this, aggregated users' interaction with the video-player using a stochastic pulse modelling process is analyzed.

In principle the concept of analyzing implicit user interaction in computing activities in order to develop user models and to provide intelligent interactions is far from new and tackles both content-based and user-based approaches. In the past, Budzik & Hammond [4] developed a system that provided dynamic responses to users based on their previous interactions with desktop computer applications. Most notably, Liu, Dolan and Pedersen [5] have improved personalization of news items by analyzing previous users' interactions with news items. In the context of multimedia, previous research has considered both content- and user-based methods for video retrieval. Kim et al. [6] report a large-scale analysis of in-video dropout and peaks in viewership and student activity, using second-by-second user interaction data from several Massive Open Online Courses (MOOCs) videos. In a different approach, Carlier, Ravindra, Charvillat and Ooi [7] propose a hybrid method where content analysis is complimented by the implicit feedback of a community of users in order to recommend viewports. Last but not least, the authors of [8] proposed a framework for analyzing human actions in video streams by introducing an implicit user-in-the-loop concept for dynamically mining semantics and annotating video streams. Based on these observations and in the current context of online multimedia sharing communities and networks, effective analysis of collective intelligence and activity remains a hot trending research topic in the area of human-computer interaction and provides motivation towards identifying new innovative research ideas.

Thus, as an example, the methodology presented herein could utilise users' interactions on content, by interpreting these interactions as time series. Such time series could be of clicks, plays or progression-slider position change of a video on YouTube<sup>1</sup>. Modelling the collective intelligence of users' interaction via the detection of characteristic patterns within the interaction time series could lead to judgment about the importance of the content, in part or whole, from which users' activity originated.

The rest of the paper is organized as follows: section II presents recent research works on implicit user interactions with multimedia content, mainly focusing on online collections. Section III provides the methodology used herein, as well as the basic steps of the proposed algorithm. Section IV has a two-fold role: initially presents the dataset used in the experimentation conducted in order to evaluate the case study results, while further discusses an evaluation procedure, which enrolled a case study in order to clearly show how the proposed mechanism works. Finally, this work ends with conclusions and thoughts as far as future work is concerned in Section V.

## II. RELATED WORK

Research concerning the temporal characteristic of human activity has recently attracted increased attention due to the increased interactivity provided by Web 2.0, as well as the increasing amounts of rich online multimedia content (such as videos) available to users through the Web.

Previous user-based research on web video has focused on the meaning of textual comments, tags, re-mixes, and micro-blogs, but has not examined simple user interactions with a web-based video player. In the seminal user-based approach to web video, Shaw and Davis [9] proposed that video representation might be better off modelled after the actual use made by users. In this way, they have analyzed the annotations to understand media semantics. Peng et al. [10] have examined the physiological behaviour (eye and head movement) of video users, in order to identify interesting key-frames. Nevertheless, the practical application of this approach is diminished as it assumes that a video camera should be available and turned-on in the home environment. Shamma, Shaw, Shafton, and Liu [11] have created summaries of broadcasts (sports and political debate, respectively) by analyzing the Twitter<sup>2</sup> stream of the respective real-time event. Although a Twitter stream is considered to be semantically rich, it lacks real-time accuracy that is required in the generation of video thumbnails, since it does require eventually a minimal amount of time (i.e. a few seconds) to type and send a text message. In contrast, the proposed method is entirely based on persistent data of real-time user interactions, such as replay button presses.

As far as research on the temporal characteristics of human social activity is concerned, several interesting areas have been identified by previous research. Issues like social behaviour analysis, social web evolution and trend detection, similar to the works of Yang and Leskovec [12], Papantoniou, Loumos, Poulizos and Rigas [13], Patrikakis, Argyriou and Papantoniou [14] and Vafopoulos [15] aim mostly in predicting social user

behaviour, in order to offer an insight to the potential customization of content. Wu and Huberman [16] and Yardi, Golder and Brzozowski [17] examined how collective human attention to items propagates and eventually fades among large populations. Given the vast amount of available online content and the ease of producing more, the authors of [18] and [19] studied the problem of predicting how much attention an item will ultimately receive. Moreover, [20] and [21] offer research on response dynamics of social systems with focus on the effects of bursts of activity in the social system. Backstrom, Kleinberg and Kumar [22] presented research on customizing feeds of news articles based on users' traffic pattern. Spatio-temporal patterns of user interaction with blog-posts have been examined by [23], [24], [25] and [26]. In addition, Aperjis, Huberman and Wu [27], presented research on online discussion forums focusing on how users behave when trying to maximize the amount of the acquired information, while minimizing the collection time.

Yang and Leskovec [11] examined temporal patterns associated with online content, by focusing on the popularity of content on social media (i.e. Twitter hashtags). Identification of the common temporal patterns is done by means of a time series clustering algorithm. Their work concludes that temporal variation of content popularity, in online social media, can be accurately described by a small set of time series shapes, while by observation of a small number of adopters of online content reliable predictions of the overall dynamics of content popularity over time can be attained.

As far as users' actions in video content is concerned and more generally, important scene selection in videos, research has mostly been based on content-based methodologies<sup>3</sup>. Nevertheless, such content-based methods often fail to capture high-level semantics that adhere to non-specialist users' navigation to videos as depicted in [28]. The authors of [29] proposed that users unintentionally show their understanding of the video content through their interaction with the viewing system. Syeda-Mahmood and Ponceleon [30] presented a client-server-based media playing and data-mining system aiming at tracking video browsing behaviour of users in order to generate fast video previews. The authors of [28], presented a user-centric approach, wherein by analysis of implicit users' interactions on web video player semantic information about the events within a video are inferred. Finally, Karydis, Avlonitis and Sioutas [31] proposed the aggregation of users' activity in order to infer important video frames.

## III. METHODOLOGY

In this section, the problem tackled in this work is formally defined and then a probabilistic algorithm solution is proposed. To begin with, let's assume a time series of users' interactions for a specific piece of content. This could be a time series of clicks or plays of a video on YouTube<sup>4</sup>, the number of times an article on a newspaper website was read, or the number of times that a hash tag in Twitter was used. The aim, therefore, is to detect patterns emerging in the temporal variation of the

<sup>1</sup> <http://www.youtube.com>

<sup>2</sup> <http://www.twitter.com>

<sup>3</sup> Interested readers may refer to [32] for an extensive survey.

<sup>4</sup> <http://www.youtube.com>

corresponding time series indicating the importance of a segment of content at a specific time interval of its duration.

The formal definition of the aforementioned scenario as a problem of time series correlation based on the correlation between the shape of the (experimentally collected) time series with the shape of a reference time series indicating local maximization of users' activity.

In the first stage, a simple procedure is used in order to average out user activity noise in the corresponding experimental signal. In the context of probability theory the noise removal can be treated, for example, with the notion of the moving average (e.g., [33]): from a curve  $s^{exp}(t)$  a new smoother curve  $s_T^{exp}(t)$  may be obtained as,

$$s_T^{exp}(t) = \frac{1}{T} \int_{t-T/2}^{t+T/2} s_T^{exp}(t') dt'$$

where  $T$  denotes the averaging "window" in time. The larger the averaging window  $T$ , the smoother the signal will be.

It must be noted that the optimum size of the averaging window  $T$  is completely defined by the variability of the initial signal. Indeed,  $T$  should be large enough in order to average out random fluctuations of the users' activities and small enough in order to reveal, and not disturb, the bell-like localized shape of the signal which in turn will demonstrate the area of high users' activity.

In the second stage, an estimation of the aggregates of users' activity via probabilistic arguments is attempted. This can be done by means of an arbitrary bell-like reference pattern: it is thus proposed that there is an aggregate of users' actions if within a specific time domain a bell-like shape of the experimental signal emerges in the sense that there is high probability that the actions are concentrated at a specific time instance (the centre of the bell) while this probability tends to zero quite symmetrically while moving away from the centre.

As a milestone of this work it is claimed that it is possible to build a scale free similarity metric, by means of the well known correlation coefficient  $(x, y)$  between two time series  $x(t)$  and  $y(t)$ , though introduction of the notion of the aforementioned reference bell-like time signal. Indeed, it can be shown that for all the cases where local maxima of the users' activity was detected, a bell like shape is encountered. Noting that the Gaussian functions are the best approximations for a bell-like shape, it is proposed that every local maxima of the popularity may be expressed as an approximated Gaussian like time signal. This assertion is depicted in Figure 1, where an arbitrary smoothed signal of density  $P(t)$  of users' activity per time instance  $t$  (in seconds) is depicted with the black continuous curve (the same interpretation is used for the rest of the Figures herein). It can be seen that within a neighbourhood of the local maxima of the time series, their shape almost perfectly matches with the upper part of a Gaussian curve (depicted by a dotted line).

However, while the Gaussian signals show high similarity in every distinct case of high popularity, the respective widths and highs are not the same in each case.

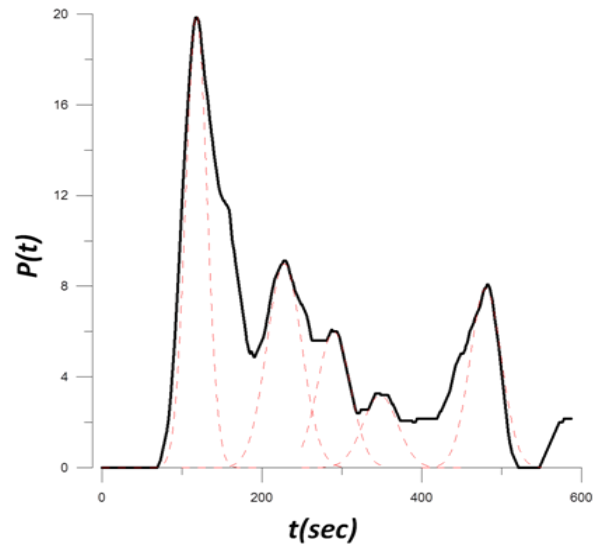


Fig. 1. The user activity signal is approximated with Gaussians bells in the neighborhood of user activity local maxima

As a result, a robust characterization of the local behaviour of high users' activities, besides the estimation of the exact location of the bell, should estimate its height and width. Moreover, the stochastic nature of users' activity signals has the direct consequence of standard signal process methods locating signal maxima and/or minima based on the estimation of first and second derivatives break down. Indeed, in real-life applications the proposed smoothing procedure is not enough in order to eliminate signal discontinuities thus resulting to infinities derivatives

#### A. Scalable stochastic similarity algorithm

In the following, a two value correlation coefficient  $r(t_c, w)$  is built, where  $t_c$  is the time centre of the Gaussian bell and  $w$  is its width. In other words, a Gaussian time signal is constructed by shifting its centre  $t_c$  over the time domain of the experimental signal and for each position a number of different Gaussian time signals are created by gradually increasing its width  $w$  (see Figure 2 from blue to red and to green solid Gaussian bell). For each Gaussian reference signal of different width the corresponding correlation coefficient is estimated with the experimental signal. The proposed formulation is motivated by the well known notion of Gaussian kernel density estimation<sup>5</sup>, a non-parametric way to estimate the probability density function of a random variable [34]. More specifically, this work further elaborates on the introduction of a series of Gaussian kernels with variable widths and finding the optimum matching or correlation coefficient for each point of the state space instead of having many Gaussians kernels with constant width and for the entire state space of the random variable under study. In this way, a two dimension correlation coefficient for each time location and for different bell widths is thus achieved. As previously stated, whenever, for a specific time centre of the Gaussian bell, a high correlation coefficient is identified during the time shifted process, a local maxima of the experimentally constructed time series is assumed. Indeed this can be seen in

<sup>5</sup> [http://en.wikipedia.org/wiki/Kernel\\_density\\_estimation](http://en.wikipedia.org/wiki/Kernel_density_estimation)

Figure 3 for an arbitrary time series (top line). The lower line, normalized to 10, depicts the corresponding correlation coefficient between the arbitrary time series and the shifted Gaussian bell. Initially the width of the bell is kept constant (as will be discussed in the corresponding section) while a robust alternative measure for the initial width could be the variance of the smoothed experimental signal).

It is evident that there is a very clear maximum of the correlation coefficient exactly when the centre of the Gaussian bell coincides with the maximum of the experimental series. As a result, the exact location of the experimental series is detected as the point of the local maximum of the corresponding correlation coefficient. Then, the assumption of the constant bell width is relaxed by keeping constant the centre of the bell by building Gaussian bells of different widths (as depicted in Figure 2). For each bell of variable width a new correlation coefficient is computed. The maximum value of this second set of correlation coefficients is estimated completing thus the process. The final result is the estimation of the maximum correlation coefficient in terms of the optimum time moment and optimum bell width. It is thus argued that the optimum time moment coincides with the local maximum of the online media popularity while the optimum Gaussian bell width coincides with the corresponding time interval over which popularity is important enough.

Summarizing, the proposed algorithm, *r-algorithm* performs the following steps: it begins with an initial Gaussian bell, the centre of which is located at the time origin of the content and its width coinciding with the variance of the smoothed experimental signal. Then, a two step procedure follows, namely the detection step and the refinement or characterization step. Within the detection step the bell is shifted along the time domain, computing the corresponding correlation coefficient between the Gaussian bell and the experimental signal.

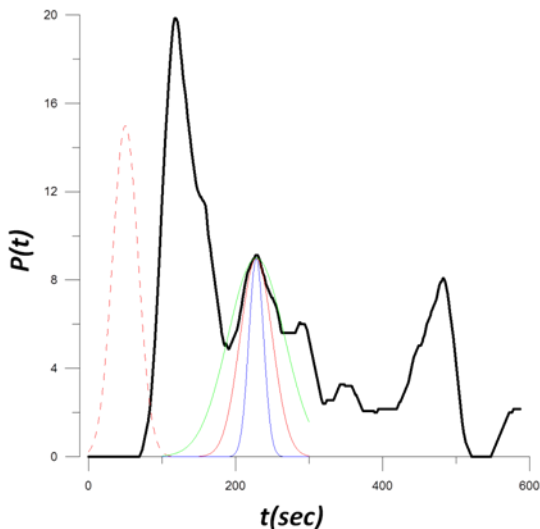


Fig. 2. The Gaussian bell is shifted over the time domain. When a local maximum of the correlation coefficient is detected a series of variables widths is created in order to estimate the optimum width

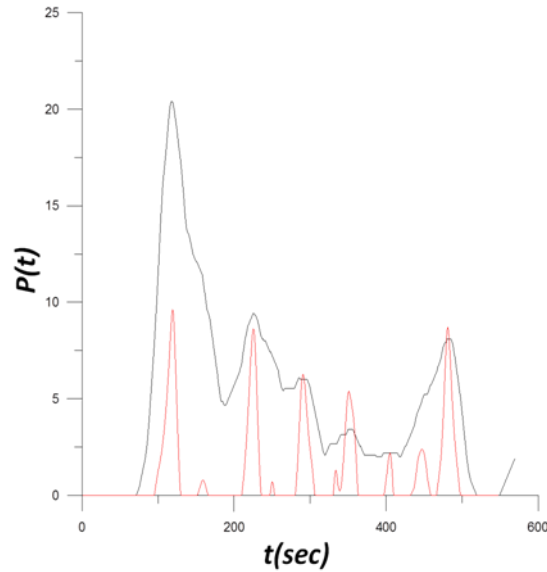


Fig. 3. Local maxima of the correlation coefficient (lower curve) coincide with local maxima of users' activity signal (upper curve)

The local maxima of users' activity are identified as the time instances where the computed correlation coefficient reaches local maximum, with the local maximum being above a user and domain specific threshold.

In the characterization step, for each local maximum of the correlation coefficient, a series of variable width Gaussian bells is generated (beginning from a value of few seconds to a fraction of the overall duration of the content) and the corresponding correlation coefficients are computed again. The calculated optimal bell width is an estimation of the interval over which the content was important enough for the users.

**Algorithm 1** The *r*-algorithm

```
Require: Experimental time series, upper part of  
Gaussian time series  $g(ct,w)$  of centre  $ct$  and  
width  $w$ .  
for  $ct=1$  to  $L$  do (detection step)  
   $r_{ct}$  (the correlation coefficient for  
  different centers)  
  if  $r_{ct} > \text{thres}$  (critical threshold of  
  correlation)  
    for  $w=1$  to  $L/10$  do (characterization step)  
       $r_{ct,w}$  (correlation coefficient for  
      variable widths)  
    end for  
  end if  
end for  
return  $r_{ct,w}$  (returns seconds of maximum users'  
activity and the corresponding time  
interval of popularity)
```

IV. ALGORITHM EVALUATION

In order to evaluate the proposed work, an open-access data-set is employed, as proposed by Gkonela & Chorianopoulos [28], which has been created in the context of a controlled user experiment, in order to ensure well-defined user-based semantics (ground-truth).

In the initialization phase, every video is considered to be associated with four distinct series in the time domain of length  $k$ , where  $k$  is the time duration of the video in seconds. Each series corresponds to the frequencies the four distinct buttons of Play/Pause, GoForward, and GoBackward are used by users at specific times. Initially, the users' activity series is created as follows: each time a user presses the GoBackward (the GoForward) button, the intervals matching the last 30 seconds (the next 30 seconds) of the video, are incremented by one, meaning that during all these 30 second the corresponding button was assumed pushed. The main experiment assumption followed relies on the fact that a user typically rewinds a video because there is something interesting to be seen again, while a user forwards a video because there is nothing of interest to see so far. In this way, a series is constructed for each button and for each video that resembles a depiction of users' activity patterns over time.

Following the above described smoothing procedure, the proposed approach focuses on the analysis of video seeking user behaviour incorporating the GoBackward and GoForward buttons. Of particular interest is the GoBackward button signal, since it may contain a quite regular pattern with a small number of regions with high users' activity. In the following, preliminary results are presented that demonstrate the proposed methodology for detecting patterns of such users' activity.

The analysis of the users' activity signals follows the implementation of the proposed *r-algorithm*. The results of the proposed methodology are depicted in Figure 4. The smoothed signal is plotted with the upper (black) curve. Moreover, the correlation coefficient of the smoothed signal with each Gaussian bell (as its centre is shifted over the time domain) is also depicted with the lower (red) curve, as extracted from the detection step of the proposed *r-algorithm*. It is evident that the exact seconds of users' activity maxima are surprisingly well estimated from the corresponding maxima of the correlation coefficient. It must be noted that while a series of small maxima is observed in Figure 4, it is obvious that the local maxima of the correlation coefficient that is aimed for must be above a very clear threshold value. This is done to avoid perturbation of the signal, which by no means expresses signal trends. For the present case a threshold value of 0.8 is assumed for the proposed algorithm. The exact threshold value depends on the problem under consideration and could be tackled by means of an establishment through a training process.

For each estimated local maxima the optimum width, as computed from the characterization step of *r-algorithm*, is also given for each detected local maxima. It can be seen in Figure 4 that the computed widths fit the corresponding widths or time interval over which important scenes are popular enough.

In order to compare the outcomes of the proposed *r-algorithm*, a surface plot is also provided in Figure 5, depicting the evolution of the two valued correlation coefficient  $r(t_c, w)$  in relation to its variables. It can be seen that the local maxima of the correlation coefficient depends on the width of the Gaussian bells is used.

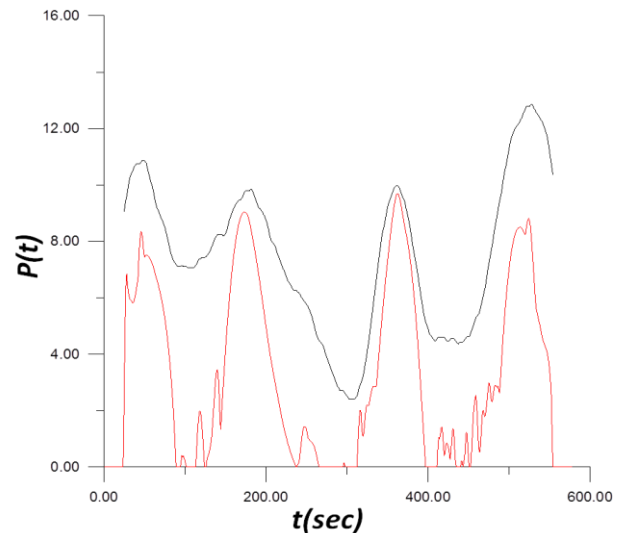


Fig. 4. Local The smoothed signal is plotted with the upper (black) curve. The optimum correlation coefficient as extracted from the detection step of the proposed *r-algorithm* is plotted with the lower (red) curve

As a result, the optimum time instance (coinciding with the local maximum of the online media popularity) and the optimum Gaussian bell width (coinciding with the corresponding time interval over which popularity is important enough) may be defined from the coordinates  $t_c, w$  of the corresponding local maxima of the plotted surface.

## V. CONCLUSIONS

In this research work describes a method that detects collective activity and identifies collective intelligence patterns via the detection of characteristic patterns within the corresponding signal monitoring aggregate activity. This framework introduces an algorithmic approach for detection of aggregates of users' activity. The latter relies heavily on the notion of a two parameter arbitrary Gaussian bell, acting as a reference pattern for aggregation. Accordingly, the aggregation of users' actions coincides to the upper part of a bell-like shape of the corresponding distribution. The complete pattern of users' interactions is defined by means of two parameters: the exact location of the centre of the Gaussian bell  $t_c$ , as well as the corresponding width  $w$ . In this way, managing to map different users' behaviour to different observed patterns is successfully achieved. As depicted herein, initial experimental research results are obtained from the application of the proposed methodology on web videos utilizing an open-access dataset. These results may be used to understand and explore social collective intelligence in online media, i.e., the way to detect users' collective behaviour, as well as how the detected collective behaviour leads to judgment about the importance of multimedia content from which users' activity originated.

Although the total improvement is not considered to be impressive, it is the belief of the authors that the approach successfully incorporates the underlying knowledge and further exploits collective activity in the video analysis value chain.



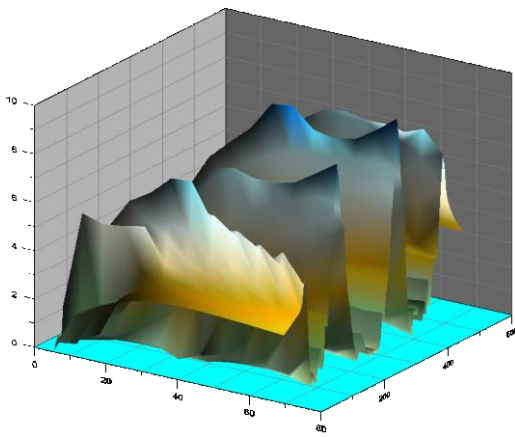


Fig. 5. A surface plot depicting the evolution of the two valued correlation coefficient  $r(t_c, w)$  in relation to its variables. Local maxima of the correlation coefficient emerge and submerge depending on the width of the Gaussian bells is used

Moreover, further research intends to use the proposed *r-algorithm* as a tool of user-based multimedia content analysis towards efficient content adaptation and personalization according to evolving users' preferences. Finally, minor enhancements on the implemented algorithmic model, e.g., in terms of additional semantic relationships exploitation, would further boost its performance and impact.

#### REFERENCES

- [1] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., Moon, S. "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," ACM SIGCOMM Conference on Internet Measurement, pp. 1-14, 2007.
- [2] Smith, J. R., Campbell, M. Naphade, M., Natsev, A., Tesic, J. "Learning and Classification of Semantic Concepts in Broadcast Video," International Conference on Intelligence Analysis, 2005.
- [3] Money, A., Agius, H. "Video summarisation: A conceptual framework and survey of the state of the art," Journal of Visual Communication and Image Representation, Vol. 19, No. 2, pp. 121-143, 2008.
- [4] Budzik, J., Hammond, K.J. "User interactions with everyday applications as context for just-in-time information access," International Conference on Intelligent User Interfaces, pp. 44-51, 2000.
- [5] Liu, J., Dolan, P., Pedersen, E.R. "Personalized news recommendation based on click behavior," International Conference on Intelligent User Interfaces, pp. 31-40, 2010.
- [6] Kim, J., Guo, P.J., Seaton, D.T., Mitros, P., Gajos, K.Z., Miller, R.C. "Understanding in-video dropouts and interaction peaks in online lecture videos," ACM conference on Learning @ scale conference, 2014.
- [7] Carlier, A., Ravindra, G., Charvillat, V., and Ooi, W. T. "Combining content-based analysis and crowd-sourcing to improve user interaction with zoomable video," ACM international conference on Multimedia, 2011.
- [8] Ntalianis, K., Doulamis, A., Tsapatsoulis, N., and Doulamis, N. "Human action analysis, annotation and modeling in video streams based on implicit user interaction," ACM workshop on Analysis and retrieval of events/actions and workflows in video streams, pp. 65-72. 2008
- [9] Shaw, R., & Davis, M. "Toward emergent representations for video," ACM International Conference on Multimedia, pp. 431-434, 2005.
- [10] Peng, W.-T., Chu, W.-T., Chang, C.-H., Chou, C.-N., Huang, W.-J., Chang, W.-Y., Hung, Y.-P. "Editing by viewing: automatic home video summarization by viewing behavior analysis," IEEE Transactions on Multimedia, Vol. 13, No. 3, pp. 539-550, 2011.
- [11] Shamma, D.A., Shaw, R., Shafton, P.L., Liu, Y. "Watch what I watch: using community activity to understand content," International Workshop on Multimedia Information Retrieval, pp. 275-284, 2007.
- [12] Yang, J., Leskovec, J. "Patterns of Temporal Variation in Online Media," ACM International Conference on Web Search and Data Mining, pp. 117-186, 2011.
- [13] Papantoniou, A., Loumos, V., Poulizos, M., Rigas, G. "A Framework for Visualizing the Web of Data: Combining DBpedia and Open APIs," Panhellenic Conference on Informatics, pp. 240-244, 2011.
- [14] Patrikakis, C.Z., Argyriou, L., Papantoniou, A. "Online Collaboration. Encyclopedia of Cyber Behavior," IGI Global, pp. 403-411, 2012.
- [15] Vafopoulos, M. "Web Science Subject Categorization," ACM WebSci, pp. 1-3, 2011.
- [16] Wu, F., Huberman, B.A. "Novelty and collective attention," National Academy of Sciences, Vol. 105, No. 45, pp. 17599-17601, 2007.
- [17] Yardi, S., Golder, A., Brzozowski, M.J. "Blogging at work and the corporate attention economy," SIGCHI Conference on Human Factors in Computing Systems, pp. 2071-2080, 2009.
- [18] Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., Hurst, M. "Patterns of Cascading Behavior in Large Blog Graphs," SIAM International Conference on Data Mining, 2007.
- [19] Szabo, G., Huberman, B.A. "Predicting the popularity of online content," ArXiv e-prints, 2008.
- [20] Barabasi, A.-L. "The origin of bursts and heavy tails in human dynamics," Nature, Vol. 435, 2005.
- [21] Crane, R., Sornette, D. "Robust dynamic classes revealed by measuring the response function of a social system," National Academy of Sciences, Vol. 105, No. 41, pp. 15649-15653, 2008.
- [22] Backstrom, L., Kleinberg, J., Kumar, R. "Optimizing web traffic via the media scheduling problem," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 89-98, 2009.
- [23] Adar, E., Zhang, L., Adamic, L.A., Lukose, R.M. "Implicit structure and the dynamics of blogspace," Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004.
- [24] Gruhl, D., Liben-Nowell, D., Guha, R.V., Tomkins, A. "Information diffusion through blogspace," International Conference on World Wide Web, pp. 491-501, 2004.
- [25] Kumar, R., Novak, J., Raghavan, P., Tomkins, A. "On the bursty evolution of blogspace," International Conference on World Wide Web, pp. 568-576, 2003.
- [26] Mei, Q., Liu, C., Su, H., Zhai, C. "A probabilistic approach to spatio-temporal theme pattern mining on weblogs," International Conference on World Wide Web, pp. 533-542, 2006.
- [27] Aperijs, C., Huberman, B.A., Wu, F. "Harvesting collective intelligence: Temporal behavior in yahoo answers," ArXiv e-prints, 2010.
- [28] Gkonela, C., Chorianopoulos, K. "Videoskip: Event detection in social web videos with an implicit user heuristic," Multimedia Tools and Applications, pp. 1-14, 2012.
- [29] Yu, B., Ma, W.-Y., Nahrstedt, K., Zhang, H.-J. "Video summarization based on user log enhanced link analysis," ACM International Conference on Multimedia, pp. 382-391, 2003.
- [30] Syeda-Mahmood, T., Ponceleon, D. "Learning video browsing behavior and its application in the generation of video previews," ACM International Conference on Multimedia, pp. 119-128, 2001.
- [31] Karydis, I., Avlonitis, M., Sioutas, S. "Collective Intelligence in Video User's Activity," Artificial Intelligence Applications and Innovations, pp. 490-499, 2012.
- [32] Geetha, P., Narayanan, V. "A Survey of Content-Based Video Retrieval," Journal of Computer Science, 4:6, pp. 474-486, 2008.
- [33] Vanmarcke, E. "Random fields, analysis and synthesis," MIT Press, 1983.
- [34] Rosenblatt, M. "Remarks on Some Nonparametric Estimates of a Density Function," The Annals of Mathematical Statistics, 27 (3):832, 1956.

# NMVSA Greedy Solution for Vertex Cover Problem

Mohammed Eshtay

Department of Computer Science  
Applied Science Private University  
Amman, Jordan

Azzam Sliet

Department of Computer Science  
University of Jordan  
Amman, Jordan

Ahmad Sharieh

Department of Computer Science  
University of Jordan  
Amman, Jordan

**Abstract**—Minimum vertex cover (MVC) is a well-known NP-Complete optimization problem. The importance of MVC in theory and practical comes from the wide range of its applications. This paper describes a polynomial time greedy algorithm to find near optimal solutions for MVC. The new algorithm NMVSA is a modification of already existed algorithm called MVAS which uses the same principle of selecting candidate from the neighborhood of the vertex with a modification in the selection procedure. A comparative study is conducted between the NMVSA and MVAS which shows that the proposed algorithm NMVSA provides better or equal results in the most cases of the underlying data sets which leads to a better average approximation ratio of NMVSA. NMVSA inherits the simplicity of the original algorithm.

**Keywords**—Vertex Cover Problem (MVC); Combinatorial Problem; NP-Complete Problem; Approximation Algorithm; Greedy algorithms; Minimum Independent Set

## I. INTRODUCTION

Many problems and concrete situations (networks, vehicle routing, maps, etc. ) are mapped into graphs and then set of algorithms are developed to manipulate these graphs to reach a solution for the given problem. Unfortunately many of such problems cannot be solved exactly in a polynomial time. Such problems are called intractable but because of their importance they must be solved.[10]

One option to solve such NP-Problems is to use heuristics. Heuristics do not guarantee the quality of solution and no guarantee also on the time required to solve the problem. Another alternative is to use approximation algorithms. In approximation algorithms we find a solution with quality between optimal and  $r$  ( $r \geq 1$ ) and we can guarantee that the algorithm will finish running in a reasonable time.

The problem that we are concern about in this paper is MVC. MVC the minimum vertex cover is a popular NP-Complete problem with a high importance in theory and practical computer science.[1].

There are many algorithms proposed to solve this problem. In this paper we will discuss set of solutions existed in the literature to solve this problem using approximation algorithms. In addition we will present new modified algorithm NMVSA and compare the results with already existed solutions of the same type.

Vertex cover is equivalent to two other popular optimization problems: MC (The Maximum Clique Problem) and MIS (Minimum Independent Set). These three problems actually can be considered as three different forms of the same problem. [5,19,20]

The following propositions hold for undirected graph  $G(V,E)$

- The minimum vertex cover does not contain vertex with degree zero.
- Every vertex cover contains  $v$  or  $w$  where  $w$  belongs to set of vertices adjacent to  $v$ , for every vertex  $v$ .
- If there is a vertex  $v$  of degree one,  $w$  is adjacent to  $u$ , there is a minimum vertex cover which contains  $w$ .
- If there is a vertex  $v$  of degree two,  $\{u,w\}$  are adjacent to  $u$  then there is a minimum vertex cover which contains both  $u$  and  $w$  or neither of them. [5,19]

**Definition:** A vertex cover in an undirected graph  $G$  is the subset of vertices  $S$  in such that every edge in the graph  $G$  is connected to at least one vertex of  $S$ . The size of a vertex cover is the number of vertices it contains. [1]

Vertex cover takes undirected graph  $G$  as input and integer number  $S$ . It tries to find a vertex cover of size  $S$ .

The size of the vertex cover is the number of vertices inside it. If the nodes are weighted by a non-negative number then it tries to find vertex cover with the minimum weighted vertices [2,15].

The vertex cover has applications in bio-informatics, security and networking. An example of vertex cover problem could be forming a team to perform certain set of tasks then we must hire enough people to accomplish each certain task. Many applications can use MVC such as network security, scheduling, biology and finance [3,16].

There are exact solutions for this problem that guarantee the optimal solution such as branch and bound but may fail to give solution within reasonable time for large instances. Other type of solution is to use heuristic algorithms. Heuristic algorithms do not guarantee the optimal solutions but they can find optimal or near optimal solution in a reasonable time [4,21,22].



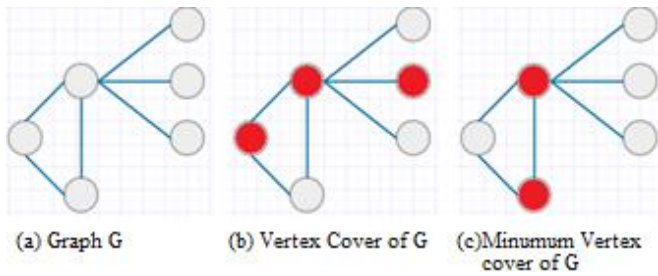


Fig. 1. Vertex Cover

### Notations

Given an undirected graph  $G(E,V)$ , a vertex cover of a graph  $G$  is a subset of vertices  $VC \subseteq V(G)$  such that every edge has an end point in  $VC$ . (That is, for all  $e \in E(G)$ ,  $e \cap VC \neq \emptyset$ ).

Fig. 1 shows examples of a vertex cover and a minimum vertex cover respectively. The nodes colored red in Fig. 1b are the nodes selected in the vector cover. The minimum vertex cover is a vertex cover of smallest possible size that appear in Fig. 1c.

The remainder of this paper is organized as follows, the different methods and related work in Section 2, the terminology, algorithm and the computation complexity in Section 3. We then discuss and analyze the results of experiments in section 4. Section 5 contains the conclusions and points to several directions for further research.

## II. RELATED WORK

The NP-Completeness of MVC problem has proved by R Karp[6]. This means that we don't have a polynomial time solution to solve it. The problem is one of the most well studied problems from researchers in computer science due to the importance and wide-range of its applications. As I mentioned earlier we can use approximation algorithms such as APPROX-VERTEX-COVER and Maximum Degree Greedy (MDG) to solve the problem.[ 7]

The Depth First Search algorithm presented by savage [8] computes the spanning tree at the beginning and then returns the non-leaves vertices of the depth first search spanning tree as a vertex cover.

The Maximum Degree Greedy (MDG) algorithm is an adaptation of a well-known heuristic algorithm used for extracting the MIS MDG keeps adding the node with the maximum degree until all edges in the graph are covered. [7,9,10]

The Greedy Independent Cover (GIC) is an adaptation of the algorithm presented in. In this algorithm we select the vertex of the minimum degree and all its neighbors to the vertex cover, the process continues until we cover all edges. [10,11]

In the algorithm LISTLEFT is proposed to find vertex cover based on a list heuristic. In this algorithm the selection of the vertices is known in advance and cannot be changes. Another list heuristic algorithm ListRight is proposed in [12].

Many of the previous mentioned methods to solve the problem of MVC depend on the degree of the vertex itself. Balaji et al. presented another technique that depends of a value called the support of the vertex. They proposed algorithm called vertex support algorithm (VSA). [6,13]

A modification of the VAS is called MVAS where the selection of the vertex does not depend only on the vertex that have the maximum the support value but it finds all the vertices with minimum support value and then it selects the vertex with minimum support from the list of all neighbors of the selected vertices. MVAS shows better results in some benchmarks comparing to original VAS. [6, 13]

Some other solutions depend on genetic algorithms such as (HVX). Xu and Ma presented solution that uses annealing algorithms to find the minimum vertex cover. In their work they show almost 100% approximation ratio for some benchmarks but they need to apply it on more benchmarks. [14].

In this paper we are introducing new greedy algorithm (NMVAS) to find the MVC by modifying MVAS. The results of applying the two algorithms in a set of benchmark sets are then compared.

## III. TERMINOLOGIES, ALGORITHM AND COMPUTATIONAL COMPLEXITY

In this section we present the proposed algorithm, the pseudo code of the algorithm, example to clarify the idea, and theoretical complexity analysis.

### A. Terminology

The selection of vertices in NMVSA algorithm relies on the degree of the neighborhood vertices. We will introduce the definition of neighborhood, degree, and support of vertex:

**Neighborhood of a vertex:** Let  $G$  be an undirected graph  $G(V, E)$  where  $V$  is set of vertices and  $E$  is set of Edges.  $|E|=m$ ,  $|V|=n$ . For each  $v \in V$  the neighborhood of  $v$   $N(v) = \{u \in V \mid u \text{ is adjacent to } v\}$ .

**Degree of a vertex:** The degree of the vertex  $d(v)$  is the number of adjacent neighbors for vertex  $v \in V$ .

**The support of a vertex:** support of a vertex  $s(v)$  is the sum of degrees of all neighbors of  $v$ .

$$s(v) = \sum_{u \in N(v)} d(u)$$

**Vertex Cover:** Vertex cover  $c = \{x \in V \mid x = u \text{ or } v \text{ if } (v,u) \text{ is an edge } e \in E\}$

### B. The proposed Algorithm

In the proposed algorithm we are trying to find an optimal or near optimal solution of the vertex cover problem because of the absence of a polynomial time solution for this NP-Complete problem. The proposed solution depends on the idea presented in VAS algorithm [17]. The selections of the vertices that will be part of the vertex cover rely on the value of support. Value of support is a value represents the sum of the degrees of the neighbors of the vertices.

The algorithm starts by finding the degree of each vertex not yet selected in the vertex cover. The degree of the vertex is the number of adjacent neighbors for vertex. The second stage after finding the degrees is finding the support value for each vertex. The algorithm proceeds by finding a list that contains all vertices that have the minimum support value. The next step is to select the vertex with the maximum support value among the neighbors of the vertices of the minimum support value. After adding the vertex to the vertex cover, all adjacent edges to this vertex are deleted. The process continues until no more edges exist.

### C. New Modified Vertex Support Algorithm (NMVSA)

The idea of selection in the proposed algorithm NMVSA shown in the algorithm 1 depends on the fact that the candidates of vertex cover are adjacent to the vertices with minimum degrees. [18] NMVSA adds a vertex from the supporting list with the maximum degree in each iteration to the vertex cover and delete all edges connected to this vertex. The process continues until no more edges still in E.

---

#### Algorithm 1: NMVSA

---

**Input:**  $G(V, E)$ .

**Output:**  $V_c$

```
1: While ( $E \neq \emptyset$ )
    //Calculate the degree for each vertex
2:   Foreach  $v_i \in V$ 
       Find  $d(v_i)$ 
3:    $\text{minSup} = \infty$ 
    //Calculate the support value for each vertex
4:   Foreach  $v_i \in V$ 
       Calculate the support of  $v_i \leftarrow s(v_i)$ 
       If  $s(v_i) < \text{minSup}$  Then
            $\text{minSup} = s(v_i)$ 
    //Create the list of minimum support vertices
5:   Foreach  $v_i \in V$ 
       If  $s(v_i) = \text{minSup}$  Then
           Add  $v_i$  to  $\text{minSupList } L$ 
    //Find the neighbors of minimum support vertices
6:   Foreach  $v_i \in L$ 
       Add  $N(v_i)$  to  $\text{neighborhoodList } H$ 
        $\text{maxSupNeighbor} = -\infty$ 
        $c = 0$  //Candidate vertex
    //Find the vertex to be added to the  $V_c$ 
7:   Foreach  $v_i \in H$ 
       If  $s(v_i) > \text{maxSupNeighbor}$  Then
            $c = (v_i)$ 
            $\text{maxSupNeighbor} = s(v_i)$ 
        $V_c = V_c \cup c$ 
        $V = V - c$ 
        $E = E - \{e \in E \mid e = (c, u), u \in V\}$ 
```

---

Step one guarantees that we pass through all edge, at each iteration we calculate the degrees of all vertices in step 2. In step 3,4 we find the support value for each vertex and we find the minimum support value among all vertices. Step 5 finds all the vertices that have the minimum support value. In step 6 we find all the neighbors of vertices listed in the minimum support vertices list. In step 7 we select the vertex with the maximum support value from the list of all neighbors of the vertices with minimum support value.

The intuition behind the algorithm is to select the vertices that connect as much as possible from the vertices that are located on the edges of the graph.

### D. Computational Complexity

According to our terminology, number of vertices is  $n$  and number of edges is  $m$ . The complexity of NMVSA can be obtained as follows: step 2 in the algorithm which is used to find the degree of each vertex requires  $O(n^2)$ . Step 4 which is used to calculate the support value of each vertex and finding the minimum support value is taking  $O(n^2)$  also. To pick all the vertices that have supporting value equal to the minimum support value in step 5 we need  $O(n)$ . Step 7 requires  $O(n)$ . The whole process will be repeated  $m$  times in the worst case.

The total running time can be expressed in the following formula:

$$m.[O(n^2) + O(n^2) + O(n) + O(n)] = O(mn^2)$$

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

The code of NMVSA has been developed using c++ on a machine Intel Core 2 Duo 2.1 GHz with 2GB memory. The algorithm has been compared to three other algorithms - VSA[17], MVSA[6], and modified MDG[2].

The programs have been executed on part of well-known dataset DIMACS benchmark set and on BHOSLIB benchmark set. The effectiveness of the algorithm NMVSA algorithm was evaluated by executing the algorithm on 40 instances of BHOSLIB 13 instances of DIMACS dataset.

Table I lists the results of all experiments hold in the algorithms. Table I consists of two parts. The first part is the length of the vertex cover found by each of the algorithms. The second part shows the ratio of the result of each algorithm and the best results obtained. The first column of the table states the name of the instance, the second column is the cardinality of the instance, and the following two columns are the results of applying MVSA and our proposed algorithm NMVSA respectively.

The second part of the table is the Approximation ratio of each of the tested algorithms.

According to Table I we can see that NMVSA gives better or equal results in all instances that we used from DIMACS benchmark. also NMVSA gives better results in 4 instances of BHOSLIB benchmark set and not as good as MVSA in two instance of BHOSLIB.

TABLE II. EXPERIMENTAL RESULTS

Benchmark	V	Optimal Vc	MVAS	NEW MVAS	Approximation ratio	
					MVAS	NEW MVAS
frb30-15-1	450	420	426	426	1.014	1.014
frb30-15-2	450	420	427	427	1.017	1.017
frb30-15-3	450	420	426	426	1.014	1.014
frb30-15-4	450	420	426	426	1.014	1.014
frb30-15-5	450	420	429	428	1.021	1.019
frb35-17-1	595	560	567	567	1.013	1.013
frb35-17-2	595	560	565	565	1.009	1.009
frb35-17-3	595	560	567	567	1.013	1.013
frb35-17-4	595	560	567	567	1.013	1.013
frb35-17-5	595	560	567	567	1.013	1.013
frb40-19-1	760	720	728	728	1.011	1.011
frb40-19-2	760	720	728	728	1.011	1.011
frb40-19-3	760	720	728	728	1.011	1.011
frb40-19-4	760	720	729	730	1.013	1.014
frb40-19-5	760	720	728	728	1.011	1.011
frb45-21-1	945	900	910	910	1.011	1.011
frb45-21-2	945	900	909	909	1.01	1.01
frb45-21-3	945	900	908	908	1.009	1.009
frb45-21-4	945	900	909	909	1.01	1.01
frb45-21-5	945	900	910	910	1.011	1.011
frb50-23-1	1150	1100	1111	1111	1.01	1.01
frb50-23-2	1150	1100	1111	1111	1.01	1.01
frb50-23-3	1150	1100	1109	1109	1.008	1.008
frb50-23-4	1150	1100	1111	1111	1.01	1.01
frb50-23-5	1150	1100	1112	1112	1.011	1.011
frb53-24-1	1272	1219	1229	1229	1.008	1.008
frb53-24-2	1272	1219	1229	1229	1.008	1.008
frb53-24-3	1272	1219	1230	1230	1.009	1.009
frb53-24-4	1272	1219	1230	1230	1.009	1.009
frb53-24-5	1272	1219	1230	1230	1.009	1.009
frb56-25-1	1400	1344	1355	1355	1.008	1.008
frb56-25-2	1400	1344	1353	1355	1.007	1.008
frb56-25-3	1400	1344	1355	1355	1.008	1.008
frb56-25-4	1400	1344	1357	1357	1.01	1.01
frb56-25-5	1400	1344	1354	1353	1.007	1.007
frb59-26-1	1534	1475	1487	1486	1.008	1.007
frb59-26-2	1534	1475	1487	1487	1.008	1.008
frb59-26-3	1534	1475	1483	1483	1.005	1.005
frb59-26-4	1534	1475	1487	1485	1.008	1.007
frb59-26-5	1534	1475	1487	1487	1.008	1.008
C500.9	500	443	451	451	1.018	1.018
C1000.9	1000	932	945	945	1.014	1.014
C2000.5	2000	1984	1988	1988	1.002	1.002
C2000.9	2000	1920	1934	1933	1.007	1.007
keller4	171	160	160	160	1	1
keller5	776	749	752	751	1.004	1.003
keller6	3361	3302	3311	3311	1.003	1.003
p_hat300-1	300	292	294	294	1.007	1.007
p_hat300-2	300	275	279	279	1.015	1.015
p_hat300-3	300	264	272	271	1.03	1.027
p_hat700-1	700	689	692	692	1.004	1.004
p_hat700-2	700	656	660	660	1.006	1.006
p_hat700-3	700	638	649	649	1.017	1.017

Table II summarizes the results of the execution of the algorithm. Table II shows that our modified version of MVSA gave equal results in 40 instances, better solutions in 7 instances and MVSA was better in two instances.

TABLE III. SUMMARY OF RESULTS

	NMVAS
Better	7
Equal	40
Worse	2

## V. CONCLUSION

In this paper we developed new algorithm called NMVSA by modifying existing algorithm called MVSA. We conducted a comparison study depending on the results of execution on two different well known benchmark sets. NMVSA gives optimal or near optimal solutions with average ratio 1.0101 on the instances that we executed the algorithm in. NMVSA generally shows better results than MVSA algorithm.

Future work includes trying to get more optimal solutions and decrease the ratio to be closer to 1

#### ACKNOWLEDGMENT

The authors are grateful to the Applied Science Private University, Amman, Jordan for the full financial support granted to this research project (Grant No. DRGS-2015-2016-54).

#### REFERENCES

- [1] Cormen, T. H. (2013). Algorithms unlocked. Mit Press.
- [2] Cai, S., Su, K., Luo, C., & Sattar, A. (2013). NuMVC: An efficient local search algorithm for minimum vertex cover. *Journal of Artificial Intelligence Research*, 687-716.
- [3] Angel, E., Campigotto, R., & Laforest, C. (2012). Implementation and comparison of heuristics for the vertex cover problem on huge graphs. In *Experimental Algorithms* (pp. 39-50). Springer Berlin Heidelberg.
- [4] Williamson, D. P., & Shmoys, D. B. (2011). The design of approximation algorithms. Cambridge university press.
- [5] Chandran, L. S., & Grandoni, F. (2005). Refined memorization for vertex cover. *Information Processing Letters*, 93(3), 125-131.
- [6] Imran, K., & Hasham, K. Modified Vertex Support Algorithm: A New approach for approximation of Minimum vertex cover. *Research Journal of Computer and Information Technology Sciences* ISSN, 2320, 6527.
- [7] Bansal, S., & Rana, A. Analysis of Various Algorithms to Solve Vertex Cover Problem.
- [8] Savage, C. (1982). Depth-first search and the vertex cover problem. *Information Processing Letters*, 14(5), 233-235.
- [9] Bodlaender, H. L., Thilikos, D. M., & Yamazaki, K. (1997). It is hard to know when greedy is good for finding independent sets. *Information Processing Letters*, 61(2), 101-106.
- [10] Delbot, F., & Laforest, C. (2010). Analytical and experimental comparison of six algorithms for the vertex cover problem. *Journal of Experimental Algorithmics (JEA)*, 15, 1-4.
- [11] Halldórsson, M. M., & Radhakrishnan, J. (1997). Greed is good: Approximating independent sets in sparse and bounded-degree graphs. *Algorithmica*, 18(1), 145-163.
- [12] Delbot, F., & Laforest, C. (2008). A better list heuristic for vertex cover. *Information Processing Letters*, 107(3), 125-127.
- [13] Khan, I., & Khan, S. (2014). Experimental Comparison of Five Approximation Algorithms for Minimum Vertex Cover. *International Journal of u-and e-Service, Science and Technology*, 7(6), 69-84.
- [14] Xu, X., & Ma, J. (2006). An efficient simulated annealing algorithm for the minimum vertex cover problem. *Neurocomputing*, 69(7), 913-916.
- [15] Hammo, B., Sleit, A., & El-Haj, M. (2007). Effectiveness of query expansion in searching the Holy Quran.
- [16] Sleit, A., Abusharkh, S., Etoom, R., & Khero, Y. (2012). An enhanced semi-blind DWT-SVD-based watermarking technique for digital images. *The Imaging Science Journal*, 60(1), 29-38.
- [17] Balaji, S., Swaminathan, V., & Kannan, K. (2010). Optimization of unweighted minimum vertex cover. *World Academy of Science, Engineering and Technology*, 43, 716-729.
- [18] Gajurel, S., & Bielefeld, R. (2012). A fast near optimal vertex cover algorithm (novca). *IJEA*, 3(1), 9-18.
- [19] Sharieh, A., Al Rawagepfeh, W., Mahafzah, M., & Al Dahamsheh, A. (2008). An algorithm for finding maximum independent set in a graph. *European Journal of Scientific Research*, 23(4), 586-596.
- [20] Al-Jaber, & A., Sharieh, A. (2000). Algorithms based on weight factors for maximum independent set, *DIRASAT, Nat. Sci.*, 27(1), 74-90.
- [21] Sleit, A., Serhan, S., & Nemir, L. (2008, August). A histogram based speaker identification technique. In *Applications of Digital Information and Web Technologies, 2008. ICADIWT 2008. First International Conference on the* (pp. 384-388). IEEE.
- [22] Sleit, A., Qatawneh, M., Al-Sharief, M., Al-Jabaly, R., & Karajeh, O. (2011). Image Clustering using Color, Texture and Shape Features. *KSII Transactions on Internet & Information Systems*, 5(1).

# Implementation of Pedestrian Dynamic In Cellular Automata Based Pattern Generation

Purba Daru Kusuma  
Department of Computer System  
Telkom University  
Bandung, Indonesia

**Abstract**—Pattern generation is one of the ways to implement computer science in art. Many methods have been implemented. One of them is cellular automata. In a previous work, cellular automata (CA) has been used to create an image with stochastic and irregular pattern. There are problems in the performance of the method because the average number of the occupied cells is less than 50 percent. So, this method must be improved. In this research, the pedestrian dynamic concept is implemented into the pattern generation process. This method is used so that there is a combination between stochastic and deterministic approaches in generating the pattern. This combination is the key element of the method. This proposed model has successfully produced irregular pattern image too. Based on quantitative test, the occupied cell ratio is still less than 50 percent but the proposed model can make a better distance between last position and starting point nodes of the pattern. When the number of agents is 75, the target to reach the occupied cells ratio by more than 75 percent is achieved.

**Keywords**—*pattern generation; cellular automata; pedestrian dynamic; intelligent agent*

## I. INTRODUCTION

Pattern generation is one of the ways to implement computer science in art. Many methods have been implemented to create the pattern automatically. Some methods used continuous approach [1,2]. Others use discrete ones. One of the popular methods that have been used is cellular automata (CA). This method has been used because of its simplicity and less complex nature. This method is suitable for generating homogenous structure pattern. One of the examples of homogenous patterns are Tuntrum and Rangrang as traditional patterns in Indonesia.

The previous work has been successfully creating images with irregular pattern by using CA [3]. This work used a fully stochastic approach. The previous research has met its goal. On the other side, there are some problems in its performance. The ratio of occupied cells is less than 50 percent. The distance between the starting point to the last point is too short. So the method needs to be improved.

In this research, the pedestrian dynamic approach is used to improve pattern generation performance. This approach is used because of its characteristics. Basically, pedestrian dynamic is a deterministic model. This approach is used so that there is a combination between deterministic and stochastic approaches in this pattern generation model.

The organization of this paper is as follows. The first section is the explanation about the previous work and the

problems so the research needs to be continued. The second section is the explanation about pedestrian dynamic concept and the reason why it can be implemented in the pattern generation research. The third section is the explanation about the proposed method so the performance can be improved, which is the combination between pedestrian dynamic method and the previous method. The fourth section is the discussion about testing for the new proposed model and the result. The fifth section is the conclusion of the analysis and the proposal for future work.

## II. PEDESTRIAN DYNAMIC

Pedestrian dynamic is one of the most interesting research areas. Many researches have done work in this area, especially in computer science. Many of these researches are multi disciplinary in nature because there are many aspects that need to be used to formulate this behavior, depending on its complexity. Apart from the physical aspect, many methods implement psychological aspect too [4,5].

Pedestrian dynamic researches have been implemented into many applications. Many of them are used in simulation, from just pedestrian traffic to crowd simulation [6]. Some of the simulations are useful in improving the evacuation process [7] or building design.

There are two approaches in pedestrian dynamic modeling: the continuous and the discrete ones. The most popular continuous method is the social forces model [4,5,7-9]. It is popular because it is precise. The disadvantage of this method is its complexity in determining the speed and the direction of the movement. The most popular discrete method is cellular automata [6,10-17]. It is popular because it is simple and light. Therefore, this method can be used in simulating the huge crowd or traffic.

There are two points of view in pedestrian dynamic modeling, the macroscopic and the microscopic ones [9]. The macroscopic one focuses on the situation of the area [9]. So, there are many simplifications and generalization. The microscopic one focuses on the individual behavior [9]. There are many aspects that may be added in the individual decision making process and movement.

In pedestrian dynamic modeling, a person will move from their starting point to the destination. When a person occupies some area, the others cannot occupy this area. Person can move only to the empty area. When a person moves to another area, the previous occupied area is now free and can be occupied by other persons.

This characteristic is useful in pattern generation. The movement path can be viewed as a pattern. The difference is in pattern generation; area that has been occupied by a person cannot be occupied by other persons. So, pedestrian dynamic can be adopted in pattern generation model.

### III. PROPOSED MODEL

In this research, CA-based pedestrian dynamic is proposed to be combined in the pattern generation model. It is because the previous work [3] is based on CA too, so they have same approach, the discrete one. In CA-based pedestrian dynamic, the occupied area is represented in cell. In this model, an intelligent agent is used to represent a person moving from a starting point to the destination. Four-neighborhood cell is implemented, so there are four possible directions for every agent in every time step.

The proposed model is the combination between deterministic and stochastic approaches. The stochastic approach is adopted from the previous work [3]. The deterministic approach is adopted from the pedestrian dynamic. There is weighting between the stochastic and the deterministic approaches.

In this proposed model, space is represented in two-dimensional arrays: x coordinate is used to represent the horizontal position and y coordinate is used to represent the vertical one. The (0,0) position is located in the left top. Variable b with its value from 1 to 4 represents the directions of agent and it represents left, top, right, and bottom consecutively. Some variables are used in this model. The model of the transition for each agent is explained in Equation 1 to 3.

n = number of agent

T = number of iteration

t = iteration at t

$a_i$  = agent with index i

$a_{i,x,y}$  = current position of the agent

$s_{i,x,y}$  = starting position of agent i

$d_{i,x,y}$  = destination position of agent i

$w_1$  = weight of deterministic part

$w_2$  = weight of stochastic part

k = random number in stochastic part

$\delta_{x,y}$  = cell status

$P_{ba_i}$  = probability in direction b for agent i

$D_{ca_i}$  = distance between agent i's current position to its destination

$D_{ba_i}$  = distance between direction b of agent i to its destination

$$\max P_{ba_i} \tag{1}$$

$$P_{ba_i} = (w_1(D_{ba_i} / D_{ca_i}) + w_2k) \delta_{x,y} \tag{2}$$

$$w_1 = 1 - w_2 \tag{3}$$

During the iteration, agent will move to its possible direction with the highest probability. The value of  $\delta$  is 1 if the cell is free and 0 if the cell has been occupied. So, if the  $\delta$  of the cell is 0, the probability to occupy this cell is 0 too. If the probability of all directions is 0, this agent cannot move again. The other conditions that agent will not move are if the iteration t has reached T or the agent has reached its destination which is represented as  $a_{i,x,y} = d_{i,x,y}$ . The illustration can be seen in Figure 1.

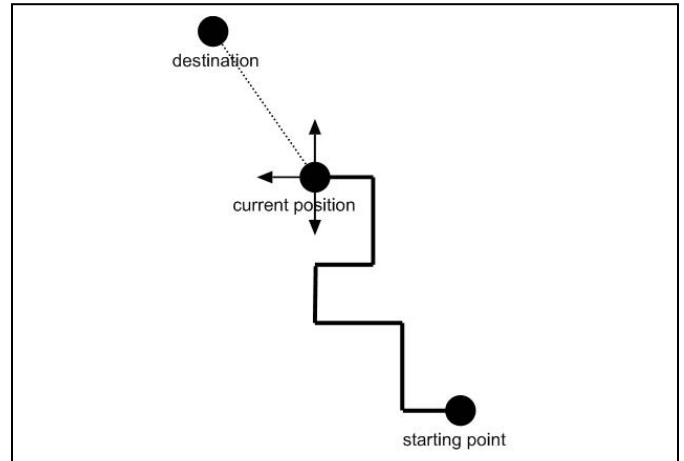


Fig. 1. Movement Path

Before the iteration starts, the starting point and the destination must be determined. In this model, the starting point and destination are generated randomly. In this process, there are two rules that are implemented. First, there are no agents with the same starting point as other agents. So, when the random process to determine the starting point for an agent, the checking process is done. When the position is owned by other agent, the random process still continues. The process for an agent stops when it finds a free starting point. Second, the destination must be as far as possible from the starting point so it will prevent the early stop. The early stop means the agent has reached its destination long before the iteration ends. So, model generates more than one possible destination for each agent. The chosen position is the longest distance between the starting point and the destination.

### IV. SIMULATION AND TESTING

Testing has been done by running the pattern generation application. The output of application is an image. Just like the previous work, there are 10 agents that create the pattern. The application iterates 100 times. The image result can be seen in Figure 2.

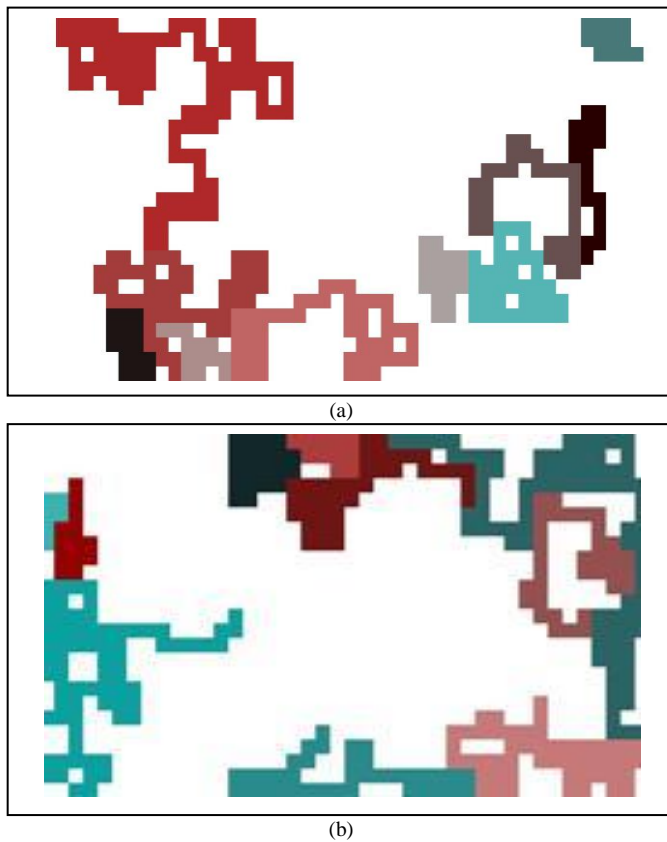


Fig. 2. Result Image

There are two images in Figure 2. Part a image is produced by fully stochastic pattern generator. Part b image is produced by deterministic-stochastic combination pattern generator. Based on visual observation, there is no difference between these two images. It's very difficult to determine which image is produced by the selected model. Thus, it can be said that the proposed model has successfully produced irregular pattern image. So, it can be continued to quantitative testing. There are three aspects that are evaluated in this paper, the occupied cells ratio, the number of cells occupied between agents, and the start-end distance.

The first testing is evaluating the occupied cell ratio. The occupied cell ratio is the ratio between occupied cells and all cells. It is presented in percent. This result can be seen in Table 1. There are 30 running sessions for this test. In this testing, the value of  $w_1$  is equal to  $w_2$ .

TABLE I. OCCUPIED CELLS RATIO

Session	Occupied Cell Ratio (%)
1	42
2	26
3	27
4	17
5	24
6	25
7	16
8	21
9	22
10	26
11	24

Session	Occupied Cell Ratio (%)
12	27
13	24
14	21
15	21
16	22
17	19
18	25
19	16
20	22
21	31
22	23
23	25
24	33
25	25
26	21
27	22
28	26
29	22
30	32
Average	24.2

Based on the data in Table 1, it can be seen that this proposed model has failed in increasing the occupied cells ratio and makes this ratio better than the result by previous stochastic model [3]. Then the testing is continued by comparing the same result by the change in the value of  $w_1$ . The result can be seen in Table 2.

TABLE II. OCCUPIED CELLS RATIO WITH INCREMENTED  $w_1$

Session	Occupied Cell Ratio (%)			
	$w_1=0.6$	$w_1=0.7$	$w_1=0.8$	$w_1=0.9$
1	25	10	38	28
2	27	27	11	41
3	23	25	34	45
4	18	36	31	27
5	26	39	35	38
6	25	26	37	32
7	27	22	35	27
8	28	31	28	41
9	22	26	33	29
10	22	39	41	32
11	27	27	21	40
12	29	27	32	44
13	20	37	23	42
14	24	36	34	38
15	16	23	37	24
16	16	32	25	29
17	22	31	38	29
18	26	28	38	28
19	22	26	36	34
20	17	31	28	34
21	26	27	23	28
22	15	22	33	34
23	18	19	41	35
24	22	19	36	34
25	26	23	21	38
26	30	34	35	26
27	28	21	28	31
28	23	28	39	36
29	29	24	41	39
30	23	28	25	37
Average	23.4	27.5	31.9	34.0

Based on the data in Table 2, it can be seen that when the weight of the deterministic part is increasing, the average of the occupied cells ratio is increasing too. Unfortunately, even the  $w_1$  value is very dominant, which is close to 1, the occupied



cells ratio is still below 50 percent. This result is still under the performance of the previous model [3]. The performance of both models is still below 50 percent. So, it is an opportunity to find another method that may increase the result so the occupied cells ratio can be higher than 50 percent.

The second testing is evaluating the performance of every agent in occupying the cells. In the first step, there are 30 running sessions with the value of  $w_1$  equal to  $w_2$ . There are three outputs, minimum occupied cells, maximum occupied cells, and average occupied cells. The result can be seen in Table 3.

TABLE III. NUMBER OF OCCUPIED CELLS BETWEEN AGENTS

Session	Number of Occupied Cells		
	min	max	average
1	4	95	51.8
2	14	55	32.5
3	4	86	32.6
4	10	46	21.7
5	17	75	30.0
6	13	68	31.1
7	6	51	20.0
8	10	49	26.6
9	13	51	27.4
10	13	78	31.9
11	5	64	30.0
12	15	83	33.7
13	4	59	29.8
14	7	62	26.3
15	10	63	26.3
16	13	85	27.3
17	12	75	23.7
18	8	100	31.1
19	1	45	19.2
20	11	65	27
21	9	100	38.2
22	9	100	29.2
23	15	62	31.8
24	17	85	40.7
25	15	100	31.4
26	7	70	25.8
27	8	48	27.6
28	18	90	32.5
29	5	49	28.3
30	11	96	39.9
Average	10.1	71.8	30.2

Based on the data in Table 3, it can be seen that the performance of the proposed model in the number of occupied cells between agents aspects is lower than the performance of the previous stochastic model [3]. This has happened in minimum, maximum, and average number of occupied cells. On the other side, the gap between the minimum and the maximum occupied cells in this proposed model is better. The test then is continued for the increment  $w_1$ . The result can be seen in Table 4.

TABLE IV. NUMBER OF OCCUPIED CELLS BETWEEN AGENTS WITH INCREMENTED  $w_1$

Session	Average Number of Occupied Cells			
	$w_1=0.6$	$w_1=0.7$	$w_1=0.8$	$w_1=0.9$
1	31.7	11.9	47.2	34.4
2	33.7	34.3	13.7	50.7
3	28.7	30.7	42.3	55.7
4	22.2	45.4	37.7	33.1

Session	Average Number of Occupied Cells			
	$w_1=0.6$	$w_1=0.7$	$w_1=0.8$	$w_1=0.9$
5	32.7	48.8	43.2	47.8
6	31.3	32.1	46.1	39.7
7	33.9	28.1	43.7	33.6
8	34.7	39.3	35.6	51.3
9	27.6	32.5	41.7	36.8
10	26.8	48.3	50.9	39
11	33.7	33.8	26.6	50
12	36.4	34.1	39.9	55.4
13	25	46.2	29.2	51.6
14	30.2	45.6	41.8	47.3
15	19.9	29.2	45.7	29.6
16	20.4	39.5	31.5	36
17	28.1	38.3	47.6	36.1
18	32.3	34.8	46.9	35.4
19	27.4	32.8	44.8	42.1
20	21.5	39	34.5	41.7
21	32.1	33.9	28.7	35.2
22	19.4	27.9	41.1	42.9
23	23	23.6	51.4	43.8
24	27.7	24	45.6	42.7
25	32.3	28.7	25.7	47.6
26	37.3	42.6	44	32.4
27	34.5	26.1	34.5	38.9
28	28.2	35.4	48.2	44.7
29	36.1	30.5	50.8	47.9
30	29.1	35	31.4	45.8
Average	29.3	34.4	39.7	42.3

Based on the data in Table 4, it can be seen that the increasing in the weight of the deterministic part increases the number of occupied cells. Unfortunately, even when the  $w_1$  is close to dominant, the performance is still below 50 percent than the maximum opportunity. This condition is correlated with the first testing.

The third testing is evaluating the distance between the starting point with the last position of the agent. There are 30 running sessions for this test. In the first step, the value of  $w_1$  is equal to  $w_2$ . The result can be seen in Table 5.

TABLE V. DISTANCE BETWEEN STARTING POINT AND LAST POSITION

Session	Distance		
	Min	max	average
1	1.4	27.3	13.3
2	3.6	19.0	12.1
3	2.0	23.4	12.5
4	3.6	23.0	10.6
5	2.0	23.1	12.8
6	2.0	23.3	11.1
7	4.5	24.0	12.6
8	2.2	24.0	12.1
9	1.4	24.0	10.1
10	1.0	24.0	14.0
11	1.0	17.1	9.0
12	1.0	23.0	13.6
13	1.0	16.3	6.2
14	2.0	25.8	12.1
15	6.3	24.0	15.6
16	1.0	21.0	10.9
17	2.2	20.6	10.0
18	3.2	24.8	11.2
19	3.0	17.7	9.8
20	5.1	16.0	10.1
21	2.8	26.2	13.3
22	2.2	22.0	11.2
23	3.2	17.5	10.6

Session	Distance		
	Min	max	average
24	4.0	21.0	9.2
25	5.1	24.7	14.2
26	1.4	26.8	11.5
27	3.6	23.0	13.6
28	7.8	21.1	15.3
29	5.4	22.2	13.5
30	3.2	21.5	9.3
Average	2.9	22.2	11.7

Based on the data in Table 5, it can be seen that the performance of the proposed model in distance aspect is better than the performance of the previous model. The average distance of the previous model is 10.5. The next step is evaluating the performance for the higher weight on deterministic part. The result can be seen in Table 6.

TABLE VI. AVERAGE DISTANCE WITH INCREMENTED  $W_1$

Session	Average Distance			
	$w_1=0.6$	$w_1=0.7$	$w_1=0.8$	$w_1=0.9$
1	10.3	9.3	14.7	10.9
2	10.6	10.7	10.1	16.5
3	10.5	13.8	13.5	12.9
4	10.2	15.5	11.9	7.7
5	12.3	12.5	11.1	12.2
6	10.3	10.7	17.4	8.5
7	8.1	6.9	11.5	12.5
8	12.2	12.8	9.15	13.8
9	10	11.8	11.7	12
10	11.2	13.3	13.6	10.1
11	10.8	10.2	9.3	10.4
12	11.9	11.6	10.3	11.5
13	11.9	12.7	10.9	16.3
14	11.5	11.4	12.8	12.6
15	9	8.6	10.3	9.2
16	9.3	13.1	12.9	13.5
17	12.8	11.7	15.3	13.9
18	13.1	11.3	13.9	11.2
19	10.1	9.2	11.2	15.8
20	9	10.3	12.8	14.4
21	8.9	11.4	11.5	15.8
22	9.9	12.6	12.3	13.2
23	10.4	9.3	10.4	11.1
24	7.7	14.1	10.5	12.1
25	11.7	11.3	8.8	10.7
26	13	18.2	10.7	11
27	8.8	9	10.8	9.2
28	9.2	11.6	13.6	12.7
29	10.7	13.3	11.6	9.9
30	11.5	14.2	9.6	14.1
Average	10.6	11.7	11.8	12.2

Based on the data in Table 6, when the weight of deterministic part is increasing, there is improvement in the average distance. Unfortunately, this performance is still far from its maximum potential. With 100 iterations, 100 cells width, and 50 cells height, the distance should be higher than now. So, it may be starting point and destination problem.

As the result in occupying cells is not increasing significantly, another method has been tried by increasing the number of agents. In the fourth test, the number of agents has been changed by 25, 50, and 75 agents. The result can be seen in Table 7.

TABLE VII. OCCUPIED CELLS RATIO WITH INCREMENTED NUMBER OF AGENTS

Session	Occupied Cells Ratio		
	$n=25$	$n=50$	$n=75$
1	47	74	77
2	57	74	82
3	52	73	80
4	56	68	81
5	49	75	78
6	59	71	79
7	49	78	82
8	61	66	83
9	48	73	80
10	54	76	83
11	54	71	81
12	54	68	79
13	42	70	76
14	52	70	77
15	49	71	84
16	56	75	75
17	54	71	83
18	53	73	81
19	58	70	80
20	62	70	78
21	55	68	81
22	50	76	80
23	62	68	80
24	50	69	83
25	53	76	78
26	47	68	82
27	50	68	80
28	56	77	81
29	54	73	82
30	47	76	85
Average	53.0	71.8	80.4

The occupied cells ratio increases significantly by increasing the number of agents. When number of agents is 50, the occupied cells ratio is below but close to 75 percent. When the number of agents is 75, the occupied cells ratio is 80.4 percent. So, the goal of improving the occupied cells ratio has been achieved.

## V. CONCLUSION AND FUTURE WORK

Based on the explanation above, the proposed model has been developed and implemented in the pattern generation application. Based on the visual result, the pattern seems irregular. Based on the quantitative testing, the performance of the proposed model is below the performance of the previous model in occupied cells aspect. When the number of agents is increasing, the occupied cells ratio is increasing too. When the number of agents is 75, the research target to improve the occupied cells ratio by more than 75 percent has been achieved. The performance of the proposed model is better than the performance of the previous model in the distance aspect.

There are many research opportunities to improve the performance of the model, especially in the 3 aspects. There is opportunity in distributing the starting point and the destination position. The key aspect is how to avoid the premature stop of the agent before the iteration finishes.

## REFERENCES

[1] M.T.Michalewicz, Plants to Ecosystems: Modelling of Biological Structures and Processes, CSIRO Publishing, 1997.

- [2] J.Knutzen, Generating Climbing Plants Using L-Systems, master thesis, Chalmers University of Technology, 2009.
- [3] P.D.Kusuma, "Traditional Fabrics Pattern Modification using Cellular Automata", unpublished.
- [4] F.Durunipar, J.Allbeck, N.Pelechano, N.Badler, "Creating Crowd Variation with the OCEAN Personality Model", Proceeding of AAMAS, Estonil, 2008, pp.1217-1220.
- [5] F.Durunipar, From Audiences to Mobs: Crowd Simulation with Psychological Factors, Thesis, Bilkent University, 2010.
- [6] S.Bandini, F.Rubagotti, G.Vizzari, K.Shimura, "Cellular Automata Model for Pedestrian and Group Dynamics", Parallel Computing Technologies, 2011, pp.125-139.
- [7] D.Helbing, I.Farkas, T.Vicsek, "Simulating Dynamical Features of Escape Panic", Nature (407), 2000, pp.487-900.
- [8] D.Helbing,P.Molnar, "Social Force Model for Pedestrian Dynamics", Physical Review, 1995, pp.4282-4286
- [9] N.Pelechano, J.Allbeck, N.Badler, Virtual Crowds: Methods, Simulation, and Control, Morgan and Claypool Publishers, 2008.
- [10] J.Djickstra, A.J.Jessun, H.J.Timmermans, "A Multi Agent Cellular Automata Model of Pedestrian Movement", Springer-Verlag, 2001, pp.173-181.
- [11] L.Fu, W.Song, W.Lv,X.Liu,S.Lo, "Multi-grid Simulation of Counter Flow Pedestrian Dynamics with Emotion Propagation", Simulation Modelling Practice and Theory, (60), 2016, pp.1-14.
- [12] H.Lklupfel, A Cellular Automaton Model for Crowd Movement and Egress Simulation, thesis, Duisburg-Essen University, 2003.
- [13] S.Sarmady, F.Haron, A.Z.Talib, "Simulating Crowd Movement Using Fine grid Cellular Automata", 12<sup>th</sup> International Conference on Computer Modelling and Simulation, 2010, pp.428-433.
- [14] S.Sarmady, Modelling and Simulation of Movement and Behaviors in Large Crowd Using Cellular Automata, thesis, University Sains Malaysia, 2008.
- [15] A.Schadscheider, "Cellular Automaton Approach to Pedestrian Dynamics-Theory", Pedestrian and Evacuation Dynamics, 2002, pp.75-86.
- [16] W.Song, Y.F.Yu, W.C.Fan, "A Cellular Automata Evacuation Model Considering Friction and Repulsion", Science in China Series E-Engineering and Material Sciende, (48), 2005, pp.403-413.
- [17] W.Song, X.Xu, B.H.Wang, S.Ni, "Simulation of Evacuation Process Using a Multi Grid Model for Pedestrian Dynamics", Physica A, (363), 2005, pp.492-500.

# Critical Path Reduction of Distributed Arithmetic Based FIR Filter

Sunita Badave

Department of Electrical and Electronics Engineering  
M.I.T, Aurangabad  
Maharashtra, India

Anjali Bhalchandra

Department of Electronics and Telecommunication  
Engineering,  
G.E.C. Aurangabad,  
Maharashtra, India

**Abstract**—Operating speed, which is reciprocal of critical path computation time, is one of the prominent design matrices of finite impulse response (FIR) filters. It is largely affected by both, system architecture as well as technique used to design arithmetic modules. A large computation time of multipliers in conventionally designed multipliers, limits the speed of system architecture. Distributed arithmetic is one of the techniques, used to provide multiplier-free multiplication in the implementation of FIR filter. However suffers from a severe limitation of exponential growth of look up table (LUT) with order of filter. An improved distributed arithmetic technique is addressed here to design for system architecture of FIR filter. In proposed technique, a single large LUT of conventional DA is replaced by number of smaller indexed LUT pages to restrict exponential growth and to reduce system access time. It also eliminates the use of adders. Selection module selects the desired value from desired page, which leads to reduce computational time of critical path. Trade off between access times of LUT pages and selection module helps to achieve minimum critical path so as to maximize the operating speed.

Implementations are targeted to Xilinx ISE, Virtex IV devices. FIR filter with 8 bit data width of input sample results are presented here. It is observed that, proposed design perform significantly faster as compared to the conventional DA and existing DA based designs.

**Keywords**—Critical Path; Multiplier less FIR filter; Distributed Arithmetic; LUT Design; Indexed LUT

## I. INTRODUCTION

Digital Signal Processing (DSP) systems are generally implemented using sequential circuits, where numbers of arithmetic modules in the longest path between any two storage elements are members of critical path. The Critical Path Computation Time (CPCT) determines the minimum feasible clock period and hence maximum allowable operating frequency of DSP system. Finite impulse response (FIR) digital filter is one of the widely used Linear Time Invariant (LTI) systems, has gained popularity in the field of digital signal processing due to its stability, linearity and ease of implementation. However, attention need to pay specifically while designing the high speed FIR filter, as CPCT is affected by both, system architecture as well as techniques used to design arithmetic modules. For such critical design of system architecture, fixed structure offered by Digital signal processor is not suitable. However, high nonrecurring engineering (NRE) costs and long development time for application specific integrated circuits (ASICs) are making field programmable

gate arrays (FPGAs) more attractive for application specific DSP solutions. FPGA also offers design flexibility to arithmetic modules then ASICs.

For an  $N^{\text{th}}$  order FIR filter, each output sample is inner product of impulse response and input vector of latest  $N$  samples [1] given in (1).

$$Y(n) = \sum_{k=0}^{N-1} A_k X_{n-k} \quad (1)$$

For critical path minimization, direct implementation of (1) is not a cost effective solution because of two reasons. First, critical path increases with the order of filter and second, multiplier is an expensive arithmetic module with respect to area and computational time. More than two decade, many researchers [2-10] have worked on various multiplierless techniques for FIR filter design. In case of constant coefficient multiplication, look-up-table (LUT) multipliers [11-13] and distributed arithmetic (DA)[14-24] are two memory based approaches found in FIR filter design. An improved distributed Arithmetic technique is addressed here to design for system architecture for FIR filter, as its operating speed is almost independent with order of filter.

In recent years Distributed Arithmetic has gained substantial popularity due to its regular structure and high throughput capability, which results in cost-effective and efficient computing structure. This technique was first introduced by Croisier [14] and further development was carried out by Peled [15] for efficient implementation of digital filters in its serial form. Apart from its several advantages; DA based structure is facing a serious limitation of exponential growth of memory with order of filter. Many researchers [16-27] have addressed this problem, while dealing with this issues. Partial or full parallel structure with two and more than two bits [16,25] has been exploited to overcome the speed limitation, inherent to bit serial DA structure. Attempts were also been made to reduce memory requirement by recasting input data in Offset Binary Coding(OBC)[16], modified OBC and LUTless DA-OBC[19], instead of normal binary coding. Yoo and Anderson [22] extended this work and proposed a hardware efficient LUTless architecture, which gradually replaces LUT requirements with multiplexer/adder pairs. However gain in area reduction is achieved at the cost of increased critical path over the conventional design. LUT decomposition or slicing of LUT, proposed in [23], is one of the ways to restrict the exponential growth of memory. Though

this technique has elucidated a problem of exponential growth of memory, involves the fact that latency and access time are the dependent parameters of level of decomposition.

As the operating speed of a filter is governed by worst case critical path, improved technique is suggested in this paper to increase the speed of operation by reducing critical path. In proposed technique, a single large LUT of conventional DA is replaced by number of smaller indexed LUT pages to restrict exponential growth and to reduce system access time. Indexing the LUT pages eliminates the use of adders of existing techniques [16,17,19,22-24].

Selection module selects the desired value from desired page, and feed the value for further computation. Trade off between access times of LUTs and selection module helps to achieve minimum critical path so as to maximize the operating speed.

In organization of the paper, section II elaborates lookup table concept of conventional DA and proposed DA structures. Critical Path Computation Time (CPCT) analysis of previous and proposed techniques is given in section III. Section IV presents the realization of proposed architecture. Initially component level access time analysis of proposed design is presented in section V, followed by comparison of operating frequency of proposed and previous techniques. Paper is ended with conclusion, in section VI.

## II. CONVENTIONAL DISTRIBUTED ARITHMETIC ALGORITHM FOR FIR IMPLEMENTATION

Distributed Arithmetic is one of the preferred methods of FIR filter implementation, as it eliminates the need of multiplier, particularly when multiplication is with constant coefficients. By this technique, sum-of-product terms in (1), can easily be transformed into addition. Let B be the word length of input samples, then, in an unsigned binary form, X(n) can be represented as:

$$X(n) = \sum_{i=0}^{B-1} x_{n,i} 2^i \quad (2)$$

where  $x_{n,i}$  is the  $i^{\text{th}}$  bit of X(n). By Substituting the value of X(n) from (2) into (1), inner product can be expressed as:

$$Y(n) = \sum_{k=0}^{N-1} A_k \sum_{i=0}^{B-1} x_{n,i} 2^i \quad (3)$$

TABLE I. CONVENTIONAL LUT DESIGN

LUT address bits				LUT contents
$x_3$	$x_2$	$x_1$	$x_0$	
0	0	0	0	0
0	0	0	1	$A_0$
0	0	1	0	$A_1$
0	0	1	1	$A_1 + A_0$
⋮	⋮	⋮	⋮	⋮
1	1	1	1	$A_3 + A_2 + A_1 + A_0$

Interchanging the sequence of summation in (3), results into:

$$Y(n) = \sum_{i=0}^{B-1} 2^i \sum_{k=0}^{N-1} A_k x_{n,i} \quad (4)$$

Further, compressed form of (4), can be expressed as:

$$Y(n) = \sum_{i=0}^{B-1} 2^i \gamma \quad (5)$$

Where,  $\gamma = A_0 x_{0,i} + A_1 x_{1,i} + \dots + A_{N-2} x_{N-2,i} + A_{N-1} x_{N-1,i}$

$x_{ni} \in \{0,1\}$

Thus (5) creates  $2^N$  possible values of  $\gamma$ . All these values can therefore be precomputed and stored in form of look up table shown in table. I. The filtering operation is performed by successively accumulating and shifting these precomputed values, based on the bit address formed by input samples, X(n). A method is proposed to choose desired size of LUT for minimum Critical Path Computation Time of LUT unit. Let N=(n+m); where n and m are arbitrary positive integers. A single large LUT size of  $2^N$ , in conventional design is converted into  $2^m$  LUT pages, each page with  $2^n$  memory locations. Applying this concept to the (5), number of terms in  $\gamma$  can be divided into two groups: n LSB terms and m MSB terms. It is represented by:

$$\gamma = (A_0 x_{0,i} + A_1 x_{1,i} + \dots + A_{n-2} x_{n-2,i} + A_{n-1} x_{n-1,i}) + (A_n x_{n,i} + \dots + A_{n+m-1} x_{n+m-1,i}) \quad (6)$$

LSB n bits, defines the size of each LUT page, however, MSB m bits defines number of LUT pages. Instead of consisting coefficient sum in conventional look up table, proposed design LUT consists of indexed-sum-of-filter-coefficients.

TABLE II. PROPOSED LUT DESIGN

n- LUT address bits				LUT contents of each page
$x_3$	$x_2$	$x_1$	$x_0$	
0	0	0	0	I + 0
0	0	0	1	I + $A_0$
0	0	1	0	I + $A_1$
0	0	1	1	I + $A_1 + A_0$
0	1	0	0	I + $A_2$
0	1	0	1	I + $A_2 + A_0$
0	1	1	0	I + $A_2 + A_1$
0	1	1	1	I + $A_2 + A_1 + A_0$
1	0	0	0	I + $A_3$
1	0	0	1	I + $A_3 + A_0$
1	0	1	0	I + $A_3 + A_1$
1	0	1	1	I + $A_3 + A_1 + A_0$
1	1	0	0	I + $A_3 + A_2$
1	1	0	1	I + $A_3 + A_2 + A_0$
1	1	1	0	I + $A_3 + A_2 + A_1$
1	1	1	1	I + $A_3 + A_2 + A_1 + A_0$

TABLE III. INDEX TERM FOR EACH LUT PAGE

Page number	m - Address Bits		Index terms 'I' for LUT pages
	$x_5$	$x_4$	
0	0	0	0
1	0	1	$A_4$
2	1	0	$A_5$
3	1	1	$A_5 + A_4$

A page selector module selects desired output from one of the LUT pages, addressed by m bits. A desired combination of n and m facilitates to select the minimum execution time of LUT page and page selector module to attain maximum operating frequency. LUT page structure of 6<sup>th</sup> order filter, for n=4 and m=2 and indexed term of each page, is elaborated in table II and table III respectively. Each LUT page contains summation of filter coefficients and index term I.

### III. CRITICAL PATH COMPUTATION TIME ANALYSIS OF PROPOSED ARCHITECTURE

In this section, CPCT analysis [13] of conventional DA [14-16], LUTless DA [19,22], sliced DA [16,17,23,24] and proposed DA based FIR filter techniques are elaborated. These designs are taken into consideration as they are found more comparable with proposed technique.

Conventional form of distributed arithmetic FIR filter given in fig.1 consists of bank of input registers, LUT unit, and accumulator/shifter unit. Apart from these hardware units, it needs control unit, which defines sequence of filter operation.

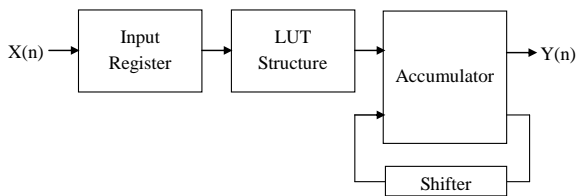


Fig. 1. Functional block diagram of conventional DA based FIR filter

Serially arriving input data values X(n) are stored in parallel form, in input register bank. Right shift of it in every clock cycle; create a word, which is used to address LUT. Successive shift and accumulation of LUT outputs in B cycle gives Y(n).

Data flow graph (DFG) of conventional DA based FIR filter, is as shown in fig.2. It consists of nodes L as LUT, A as accumulator and S as shifter. Access times of L and A are C<sub>L</sub> and C<sub>as</sub> respectively, contributes in critical path. Thus CPCT of conventional DA based FIR filter is expressed as:

$$CPCT_{(cnv)} = C_L + C_{as} \quad (7)$$

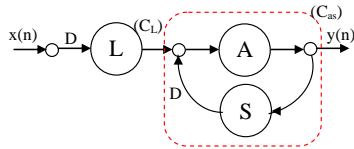


Fig. 2. Data flow graph of conventional DA based FIR filter

#### A. LUTless DA based FIR filter

Exponential growth of LUT is key issue while designing DA based FIR filter. Elimination of LUT is an attempt found in [13,24] to overcome exponential growth of LUT. In such LUTless structure, shown in fig.3, LUT is replaced by multiplexer-adder pair. On-line data generated by multiplexers are accumulated to create the filter output.

DFG of LUTless DA based FIR filter, shown in fig.4, consists of multiplexer node M, adder nodes T<sub>a</sub> and

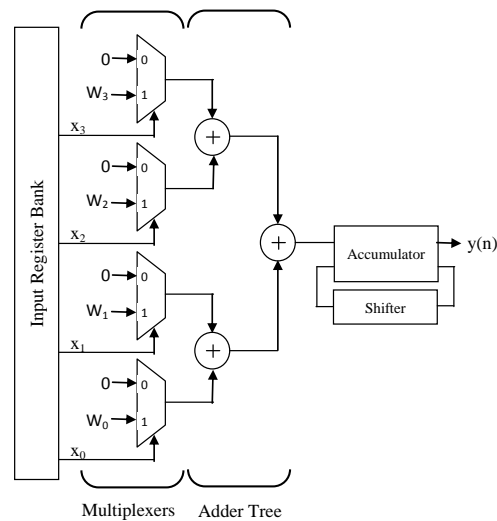


Fig. 3. LUTless DA based FIR filter

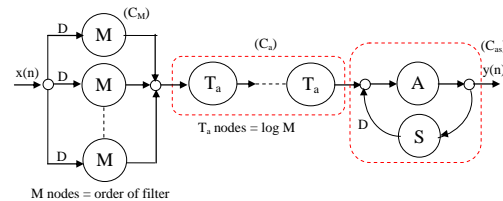


Fig. 4. Data flow graph of LUTless DA based FIR filter

accumulator node A. Though the number of multiplexers is governed by order of filter, access time of only one multiplexer contributes in CPCT, as they are operating concurrently.

Assuming the adders in adder tree are arranged in 4:2 form, access time of log<sub>2</sub>(N) adders are taken into consideration while calculating CPCT of structure C<sub>a</sub>. It will be expressed as:

$$C_a = \log_2 N \times T_a \quad (8)$$

Thus C<sub>a</sub> is highly filter order dependent as indicated in (9). CPCT of structure becomes:

$$CPCT_{(LUTless)} = C_M + C_a + C_{as} \quad (9)$$

where C<sub>M</sub> - access time of multiplexer.

C<sub>a</sub> - access time adder tree

C<sub>as</sub> - access time of accumulator/shifter unit.

#### B. Sliced LUT DA based FIR filter

Another well-known attempt found in [21,22,27] to restrict the exponential growth of LUT, is the use of multiple memory banks.

Latest, Longa and Miri [23], highlighted that, FIR filter structure will be an area efficient structure by replacing a single large LUT by number of 4-input, smaller LUTs. However, this arrangement leads to put a burden of an adder tree, as it is required to add partial terms generated by each smaller LUT. Generally such LUT arrangement is referred as partitioning or slicing of LUT. Architectural details of sliced DA based FIR filter is shown in fig.5.

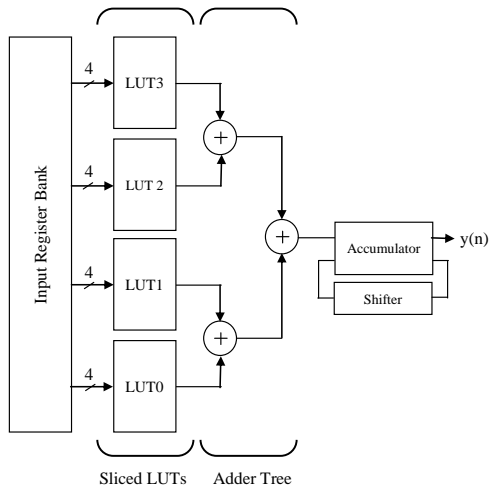


Fig. 5. Sliced LUT DA based FIR filter

Data flow graph of sliced LUT DA based FIR filter, shown in fig.6, consists of concurrently operating 4-input LUT nodes  $L_s$ , adder nodes  $T_a$ , accumulator  $A$  and shifter node  $S$ .

In this architecture, requirement of adders in adder tree is governed by number of slices. Assuming the order of filter is divisible by 4, for  $N^{\text{th}}$  order FIR filter,  $N/4$  will be number of slices and  $(N/4)-1$  will be the number of adders. Thus LUT node  $L_s$ ,  $\lceil \log_2(N/4) \rceil$  adders and accumulator are the members of critical path. So the CPCT of the structure will be:

$$CPCT_{(Slice)} = (C_{SL} + C_a + C_{as}) \quad (10)$$

Where

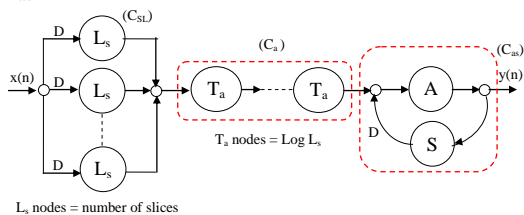
$C_{SL}$  = access time of one slice of LUT.

$C_a$  = access time of adder tree

=  $\lceil \log_2(N/4) \rceil T_a$

$T_a$  = access time of an adder.

$C_{as}$  = access time of accumulator/shifter



$L_s$  nodes = number of slices

Fig. 6. Data flow graph of sliced LUT DA based FIR filter

Access time of LUT get reduced from  $C_L$  to  $C_{SL}$  due to slicing technique, however it has added the over heads of adder tree access time  $C_a$  in  $CPCT_{(slice)}$ .

### C. Indexed LUT DA based FIR filter

LUTless and SlicedLUT has restricted the exponential growth [22,23], however it has increased the burden of access time of adder tree.

So an attempt is made, to eliminate the use of adder tree by designing an indexed LUT based FIR filter technique. In proposed design of Indexed LUT (ILUT) DA structure, node  $L$  of fig.2 is replaced by smaller, desirably indexed LUTs  $L_i$  and multiplexer  $M$ .

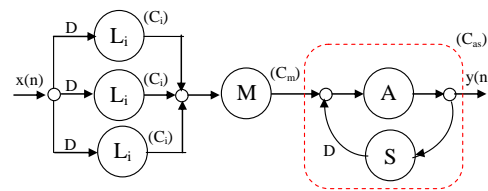


Fig. 7. Data flow graph of indexed LUT DA based FIR filter

DFG of the proposed design derived from (6), is shown in fig.7. CPCT of this structure, contributed by  $L_i$ - $M$ - $A$  nodes, will now be:

$$CPCT_{(Index)} = C_i + C_m + C_{as} \quad (11)$$

Where

$C_i$  = access time of an indexed LUT.

$C_m$  = access time of multiplexer

$C_{as}$  = access time of accumulator/shifter

Access time  $C_i$  and  $C_m$  are interdependent. The trade off of an exponentially varying LUT with linearly varying multiplexer size helps to choose optimum CPCT of a structure. Hence, improves overall operating frequency of filter. It also eliminates the need of adder tree, which further helps to improve the operating frequency.

## IV. REALIZATION OF PROPOSED ARCHITECTURE

Proposed structure of indexed LUT DA based FIR filter is elaborated in following sections. It is built up with four major components bank of input registers, look-up-table unit, accumulator/shifter unit and control unit.

### A. Input register bank

Register Bank, shown in fig.8, built up with  $N$  serial-in parallel-out shift registers, accepts  $X(n)$  input samples,  $n=0,1,\dots,N-1$ . In every clock pulse, register contents take a right shift and generates  $B$  terms of length  $N$ .

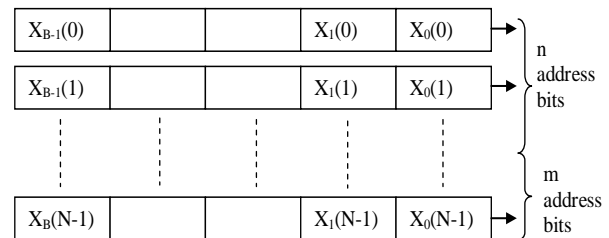


Fig. 8. Input register bank and address bifurcation

LUT address generated by register bank is split into two address groups  $n$  and  $m$ . LSB  $n$  bits define address of LUT, whereas number of LUT pages is defined by  $m$  bits.

### B. Proposed LUT unit

Indexed LUT DA based FIR filter, comprises of indexed LUT pages, each of size  $2^n$  and  $m$  bit multiplexer unit as a page selection module. It selects the desired value from desired page. Structural details of an example, considered in section 2A, of 6<sup>th</sup> order FIR filter, with  $n=4$  and  $m=2$ , is shown in fig.9. Four LUT pages, each with 16 locations are connected in parallel, by set of 4 address lines. A multiplexer unit of size 4:1 selects an appropriate output for further stage.



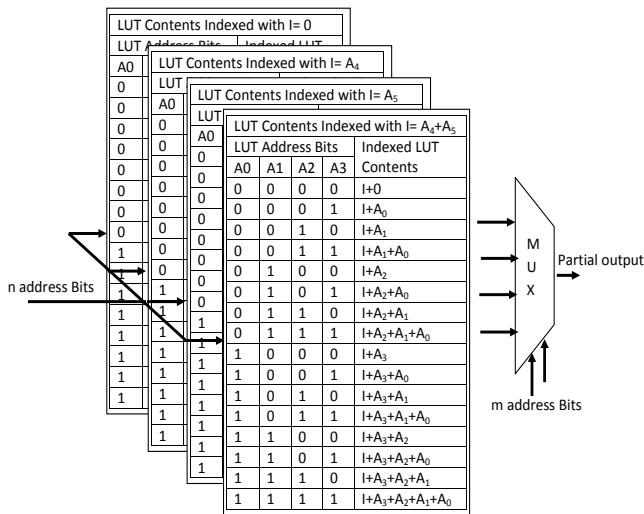


Fig. 9. Proposed structure of LUT unit

C. Accumulator and Shifter Unit

Accumulator and shifter are two separate combinational units, however jointly these are responsible for calculating the dot product term of filter output. Its hardware complexity is greatly influenced by the way of LUT addressed and accordingly a shift is given to accumulator/shifter unit to generate partial products.

D. Control Unit

It is a finite state machine, shown in fig.10, defines sequence of operation and has overall control on filtering operation.

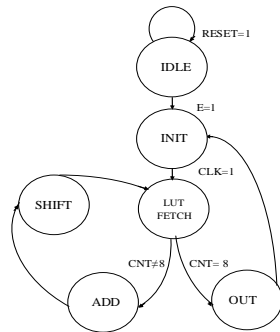


Fig. 10. Control unit of proposed structure

Filtering operation remains in idle state with application of reset. It starts with enable signal E and takes iteration equal to input precision for every clock cycle. At the end of count it gives filter output and operation begins with next fetch cycle.

V. PERFORMANCE ANALYSIS

Performance is evaluated based on operating frequency. Design is implemented on FPGA Vertex IV, for particular filter order N and for all possible combinations of n and m, as shown in table IV. Each node of proposed structure is critically analyzed for CPCT of proposed structure, for the range of filter from 4 to 8. Table IV gives the details of filter operating frequency with variation in access times of LUT page  $C_i$  and multiplexer unit  $C_m$ .

Graphical representation for 8<sup>th</sup> order FIR filter is shown in fig. 11. It indicates that, access time of LUT page  $C_i$  increases exponentially with n, at the same time access time of multiplexer  $C_m$  decreases linearly.

If  $f_{max}$  is assumed to be the maximum operating frequency,  $T_{sample}$  is the minimum time required to process each output sample, then

$$T_{sample} \geq CPCT \geq C_i + C_m + C_{as} \tag{12}$$

As

$$f_{max} = 1/T_{sample} \tag{13}$$

$$f_{max} \leq 1/C_i + C_m + C_{as}$$

As CPCT minima of filter is obtained at the point of intersection of LUT access time  $C_i$  and MUX access time  $C_m$ , which leads to maximum operating frequency. Thus filter design corresponds to these values of m and n will be treated as optimized design.

TABLE IV. ACCESS TIME ANALYSIS OF LUT UNIT MODULES

Order of Filter	Address Line distribution		LUT Unit Access time analysis		Operating freq. in MHz
	n	m	$C_i$	$C_m$	
8	7	1	6.58	3.6	151.389
	6	2	5.45	4.06	160.937
	5	3	5.02	4.46	155.876
	4	4	4.65	4.8	184.834
	3	5	4.65	5.16	169.544
	2	6	4.6	5.5	168.714
	1	7	3.84	6.1	176.625
7	6	1	5.45	3.6	189.92
	5	2	5.02	4.06	180.874
	4	3	4.65	4.46	183.441
	3	4	4.65	4.8	191.18
	2	5	4.6	5.16	182.45
6	5	1	5.02	3.6	190.13
	4	2	4.65	4.06	192.417
	3	3	4.65	4.46	205.495
	2	4	4.6	4.8	190.389
5	4	1	5.02	3.6	192
	3	2	4.65	4.06	206.793
	2	3	4.6	4.46	228.645
4	3	1	4.65	3.6	215.736
	2	2	4.6	4.06	242.93
1	3	3.84	4.46	244.09	

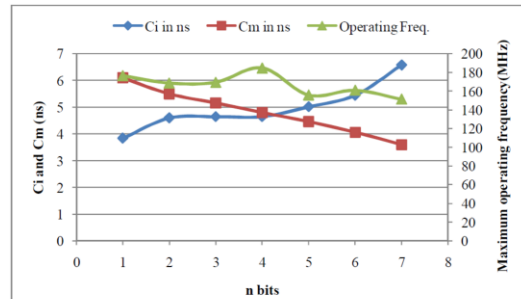


Fig. 11. Relation between access time analysis of LUT unit modules and operating frequency of 8<sup>th</sup> order FIR filter

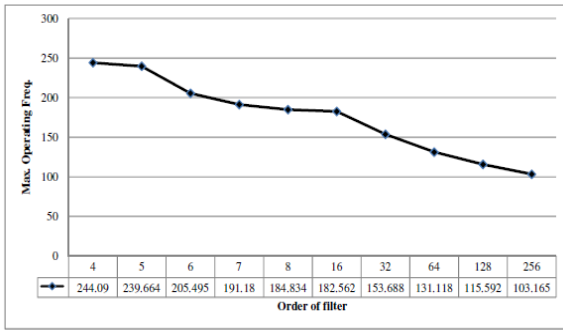


Fig. 12. Relation of maximum operating frequency with order of filter

This technique can further be extended to any desired order of filter. Filter performance upto 256 order is shown in fig. 12. Results obtained by the proposed technique are compared with Conventional DA, LUTless DA[22] and Sliced LUT DA[23]

TABLE V. OPERATING FREQUENCY COMPARISON OF VARIOUS ARCHITECTURES

Order of filter	Operating frequency of DA based filter in MHz			
	Conventional DA	LUTless DA	Sliced DA	Proposed DA design
4	242.4	242.93	240.13	244.09
5	239.01	239.06	220.037	239.664
6	200.95	174.074	200.122	205.495
7	184.65	175.503	185.685	191.18
8	176.22	174.28	167.726	184.834

techniques, which were implemented on Altera Stratix FPGA chip. To surmount the platform differences, these techniques are faithfully implemented on same platform as that of the proposed technique. Desired filter coefficients are obtained from FDATool, a special toolbox of MATLAB, which are truncated and scaled to 8-bit precision. Xilinx Integrated Software Environment (ISE) is used for performing synthesis and implementation of the designs.

To validate the correct functionality using random input, each implementation is simulated with the simulation tool provided by Xilinx.

A comparative study of maximum operating speed of conventional DA, LUTless DA, Sliced DA and proposed DA based filter techniques is presented in table V and its graphical representation is in fig.13.

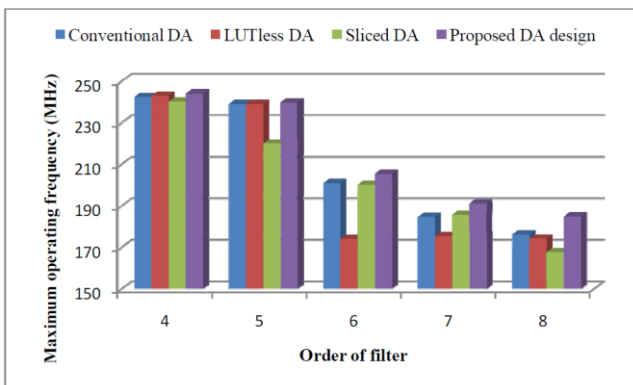


Fig. 13. Comparison of operating frequency

TABLE VI. STRUCTURAL COMPLEXITY OF PREVIOUS AND PROPOSED DESIGNS

Order of filter	Structural Complexities			
	Conventional DA	LUTless DA	Sliced DA	Proposed DA design
Input Register	$N \times B$	$N \times B$	$N \times B$	$N \times B$
Memory Bits	$M_C = 2^N \times B$	-	$M_S = (a \times 2^l) \times B$	$M_I = (2^m \times 2^n) \times B$
Decoder	$N: 2^N$	-	$a(l:2^l)$	$2^m (n:2^n)$
Number of Adders	-	$N-1$	$a-1$	-
Depth of Adders	-	$B + \log_2 N$	$B + \log_2 a$	-
Multiplexers	-	-	-	$2^m:1$
CPCT	$C_L + C_{as}$	$C_M + C_a + C_{as}$	$C_{SL} + C_a + C_{as}$	$C_i + C_m + C_{as}$
Latency	$B+1$	$B+1$	$B+1$	$B+1$
Throughput	$B+2$	$B+2$	$B+2$	$B+2$

Operating frequency reduces with the order of filter is one of the obvious observations indicated in table V. It is also observed that operating frequency of proposed technique is higher than conventional DA and existing DA[22,23] techniques. No much gain in rise of frequency is obtained at 4<sup>th</sup> order as techniques are get correlated with technology platform, however frequency growth is increasing along with the order of filter.

Structural complexities of N<sup>th</sup> order filter are analyzed and performances are compared for random input samples x(n). Word length of input sample and filter coefficient is assumed to be of B bits, which makes size of input register bank to be same for all designs under consideration. Latency and throughput found same in all DA based structures; however operating speed of individual technique makes the value to differ.

For implementation of N<sup>th</sup> order conventional DA based FIR filter requires memory array of  $2^N \times B$  bits and the size of decoder is  $N:2^N$ . CPCT of the structure is  $(C_L + C_{as})$ , increases exponentially due to exponential rise in  $C_L$ , however  $C_{as}$  is independent with order of filter. Thus it is almost constant in all structures. Structural complexities of conventional DA based FIR filters are considered as bench marks for performance comparison.

Slicing of single large memory reduces the memory requirement of design from  $2^N \times B$  of conventional DA to  $(a \times 2^l) \times B$ ; where a and l are the factors of N. Thus decoder also get changed from single  $N: 2^N$  to  $a, l:2^l$ . As multiple terms are generated by this technique, need at least a-1 adders to generate coefficient sum as partial term. A single large LUT is replaced by smaller LUTs, reduces LUT access time from  $C_L$  to  $C_{SL}$ , however it adds adder access time  $C_a$ , tending to increase CPCT of structure.

LUTless technique selects filter coefficient on-line by multiplexer, eliminates the need of memory and corresponding decoder at the cost of N-1 adders. As LUT is replaced by multiplexers and adders,  $C_M$  and  $C_a$  are the contributors of CPCT, which are highly filter order dependent.

In proposed technique, indexing of LUT pages reduces its access time  $C_i$  instead of  $C_L$  as well as eliminates  $C_a$  as a prime contributor of CPCT of LUTless and sliced LUT DA based techniques. It adds a small burden of LUT page selection

module Cm, to CPCT of structure. However it leads to reduce overall CPCT, leading to increase in operating frequency. This rise in frequency is significant with higher filter order as indicated in table V.

## VI. CONCLUSION

For high speed FIR filter implementation in distributed arithmetic, the exponential rise of memory access time with the filter coefficients has always been considered to be a fundamental drawback. LUTless DA and sliced LUT DA based technique restricts exponential growth, however needs adders to generate partial term. Number of adders and depth of adders, is governed by order of filter in LUTless technique. However in sliced LUT based technique, number of slices defines number of adders. Even for particular filter order, number of adders increases with increase in number of slices, tending to increase CPCT of structure. An innovative technique to reduce CPCT of FIR filter is designed and implemented successfully, which leads to increase in operating frequency. Indexing of LUT restricts exponential growth and also completely eliminates need of adders which results in significant reduction in CPCT and maximizes operating frequency.

## REFERENCES

- [1] Mitra S. K., Digital filter structures: Digital Signal Processing-A Computer Based Approach, 3<sup>rd</sup> ed., India.Tata McGraw Hill, 2008, pp.427-437
- [2] Henry Samueli, "An Improved Search Algorithm for the Design of Multiplierless FIR Filters with Powers-of-Two Coefficients", IEEE Transactions on Circuits and Systems, Vol 3.6 , No.7, pp.1044-1047, July1989.
- [3] Joseph B. Evans, "An Efficient FIR Filter Architectures Suitable for FPGA Implementation", Proceedings of the IEEE International Symposium on Circuits and Systems(ISCAS), pp.226-228, 1993.
- [4] Woo Jin Oh, Yong Hoon Lee, "Implementation of Programmable Multiplierless FIR Filters with Powers-Of-Two Coefficients", IEEE Transactions on Circuits and Systems-II Analog and Digital Signal Processing, Vol. 42, No. 8, pp.553-556, August 1995.
- [5] Kei-Yong Khoo, Alan Kwentus, and Alan N. Willson, Jr., "A Programmable FIR Digital Filter Using CSD Coefficients", IEEE Journal of Solid-State Circuits Vol II No 6, pp.869-874, June 1996.
- [6] Dawoud D. S., "Realization of pipelined multiplier - free FIR digital filter", Proc. IEEE Africon Conference, pp.335-338, 1999.
- [7] Marko Kosunen, Kari Halonen, "A Programmable Fir Filter Using Serial-In-Time Multiplication And Canonic Signed Digit Coefficients" Proceedings of the 7<sup>th</sup> IEEE International Conference on Electronics, Circuits and Systems(ICECS), pp.563-566, 2000.
- [8] Kah-Howe Tan, Wen Fung Leong, Kadambari Kaluri, M A. Soderstrand and Louis G. Johnson, "FIR Filter Design Program that Matches Specifications Rather than Filter Coefficients Results in Large Savings in FPGA Resources" Proceedings of the IEEE International Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems and Computers, Vol.2, pp.1349-1352, 2001.
- [9] Zhangwen Tang, Jie Zhang and Hao Min, "A High-Speed, Programmable, CSD Coefficient FIR Filter", IEEE Transactions on Consumer Electronics, Vol. 48, No. 4, pp. 834-837, November 2002.
- [10] K. S. Yeung and S. C. Chan, "Multiplier-Less Digital filters Using Programmable Sum-of-Power-of-Two(SO POT) Coefficients", Proceedings of the IEEE International Conference on Field-Programmable Technology(FPT), pp.78-84, 2002.
- [11] Pramod Kumar Meher, "New Look-up-Table Optimizations for Memory-Based Multiplication", Proceedings of the 12<sup>th</sup> IEEE International Symposium on Integrated Circuits, ISIC 2009, pp.663-666, 2009.
- [12] Pramod Kumar Meher, "New Approach to Look-Up-Table Design and Memory-Based Realization of FIR Digital Filter", IEEE Transactions on Circuits and Systems—I: Regular Papers, Vol. 57, No. 3, pp. 512-603, March 2010.
- [13] Keshab K. Parhi, "VLSI Digital Signal Processing- Design and Implementation," in John Wiley & Sons, India,1999, pp.36-37,43-45.
- [14] A. Croisier D. J. Esteban, M. E. Levilion, and V. Rizo, "Digital filter for PCM encoded signals," *U.S. Patent* 3 777 130, Dec. 4,1973
- [15] A. Peled and B. Liu, "A new hardware realization of digital filters," in IEEE Trans. Acoust. Speech, Signal Process., vol. 22, no. 6, pp. 456-462, Dec.1974.
- [16] S.A.White, "Applications of the Distributed Arithmetic to Digital Signal Processing: A Tutorial Review," IEEE ASSP Mag., vol. 6, no. 3, pp. 5-19, Jul. 1989.
- [17] Wayne P. Burleson, Louis L. Scharf, "VLSI Design of Inner-Product Computers Using Distributed Arithmetic", Proceedings of the IEEE International Symposium on Circuits and Systems(ISCAS), pp.158-161, 1989.
- [18] Rudi BabiE, Mitja Solar, Bruno Stiglic, "High Order FIR Digital Filter Realization in Distributed Arithmetic" Proceedings of 6th Mediterranean Electrotechnical Conference, pp.367-370,1991.
- [19] Jung-Pal Choi Seung- Cheol Shin Jin- Gyun Chung, "Efficient Rom Size Reduction For Distributed Arithmetic", IEEE International Symposium on Circuits and Systems(ISCAS), pp. II-61 to II-64, 2000.
- [20] T. S.Chang and C.-W.Jen, "Hardware-efficient pipelined programmable FIR filter design", Proceeding of IEEE on Computers and Digital Techniques, vol.148, issue: 6, pp.227-232,2001.
- [21] Chin-Chao Chen, Tay-Jyi Lin, Chih-Wei Liu, and Chein-Wei Jen, "Complexity-Aware Design of DA-Based FIR Filters", proceeding of IEEE Asia-Pacific Conference on Circuits and Systems, pp.445-448,2004.
- [22] Heejong Yoo And David V. Anderson, "Hardware-Efficient Distributed Arithmetic Architecture For High-Order Digital Filters" Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP),Vol.5, pp.V-125-V-128, 2005.
- [23] Patrick Longa And Ali Miri, "Area-Efficient FIR Filter Design On FPGA Using Distributed Arithmetic", IEEE International Symposium on Signal Processing And Information Technology, pp. 248-252, 2006.
- [24] M. Mehendale, S. D. Sherlekar, and G.Venkatesh, "Area-delay Tradeoff in distributed arithmetic based implementation of FIR filters," Proceedings of 10<sup>th</sup> International Conference of VLSI Design, pp. 124-129, 1997.
- [25] Shiann-Shiun Jeng, Hsing-Chen Lin, And Shu-Ming Chang, "FPGA Implementation of FIR Filter Using M-Bit Parallel Distributed Arithmetic" IEEE, International Symposium on Circuits and Systems(ISCAS),pp.875-878,2006.
- [26] Pramod Kumar Meher, Shrutisagar Chandrasekaran, Abbes Amira, "FPGA Realization of FIR Filters by Efficient and Flexible Systolization Using Distributed Arithmetic", IEEE Transactions on Signal Processing, Vol. 56, No. 7, pp.3009-3017, 2008.
- [27] B.K.Mohanty,P.K Meher, "A High-Performance FIR Filter Architecture for Fixed and Reconfigurable Applications", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, issue 99,1-9, 2015.

## AUTHOR PROFILES



**Sunita Mukund Badave** received the B.E. degree in Electrical (Electronics Specialization) from Shivaji University in 1989 and M.E.Degree in Electrical from Dr.B.A.M.University., Aurangabad, India, in 1998. She is currently working toward the Ph.D. degree in Electronics at Dr.B.A.M.University. Her research interests include architectures and circuit design for digital signal processing. She has presented nearly 16 technical papers in at Nationally and Internationally. Mrs. S.M. Badave is *Member* of the Institute of Electronics and Telecommunication Engineers(IETE),India and life member of Indian Society for Technical Education(ISTE)India.She is also member of IAENG, International Association of Engineers.



**Anjali S. Bhalchandra** received the B.E. Electronics and Telecommunication degree and M.E. Electronics degree in 1985 and 1992 respectively. She has completed her Ph.D.

in Electronics from S.R.M.University, Nanded, India, in 2004. She has a scientific and technical background covering the areas of Electronics and Communication. Currently, she is Head of Electronics and Telecommunication Engineering Department and Associate Professor in Government College of Engineering, Aurangabad. Her research interest includes image processing,

signal processing and communication. She has published more than 50 technical papers in various reputed journals and conference proceedings. Dr. Bhalchandra is a Fellow of the Institution of Engineers (IE), India and life member of Indian Society for Technical Education(ISTE)India.

# Explorative Study of SQL Injection Attacks and Mechanisms to Secure Web Application Database- A Review

Chandershekhar Sharma

Research Scholar, Computer Science  
and Engineering  
Rajasthan Technical University  
Kota, India

Dr. S. C. Jain

Professor, Computer Science and  
Engineering  
Rajasthan Technical University  
Kota, India

Dr. Arvind K Sharma

Department of CSI  
Kota University  
Kota, India

**Abstract**—The increasing innovations in web development technologies direct the augmentation of user friendly web applications. With activities like - online banking, shopping, booking, trading etc. these applications have become an integral part of everyone's daily routine. The profit driven online business industry has also acknowledged this growth because a thriving application provides the global platform to an organization. Database of web application is the most valuable asset which stores sensitive information of an individual and of an organization. SQLIA is the topmost threat as it targets the database on web application. It allows the attacker to gain control over the application ensuing financial fraud, leak of confidential data and even deleting the database. The exhaustive survey of SQL injection attacks presented in this paper is based on empirical analysis. This comprises the deployment of injection mechanism for each attack with respective types on various websites, dummy databases and web applications. The paramount security mechanism for web application database is also discussed to mitigate SQL injection attacks.

**Keywords**—Injection Attacks; SQL vulnerabilities; Web Application Attacks

## I. INTRODUCTION

Rapid advancement in web technologies has expedited the rate of adoption of database driven web application. The backend database servers of these web applications accumulate some general data along with critical & sensitive information about organizations and clients [40].As the database is accessible from anywhere over internet makes it prone to attacks. The most hazardous attacks against database driven web applications are –SQL injection attacks [48]. These attacks are very serious threat to any web application that receives inputs from user and incorporate it in to SQL queries to an underlying database[10][38].Although web application keeps the user's data secure for making any online exchange of information but presence of vulnerabilities makes this attack feasible. SQLIA are mostly caused by the insufficient validation of user input.

An attacker can submit a query (utilizing SQL command) directly to the database which can extract categorical information depending upon the severity of vulnerability. Database is the main asset of any web application to which attackers are keenly fascinated. SQL injection is very lucrative

for attackers as there is a successful black market that deal all scarcely digitally purloined data like credit card information, bank accounts detail and social security numbers etc.[21].

With a little knowledge of SQL commands and ingenious conjecture work to crucial table name SQL injection attacks can be launched. These commands alter the desired output of queries to break in to database. Injection attacks were ranked 1st attack in 2013 by OWASP (Open Web Application Security projects) in TOP ten attacks and found that 80% of web applications are Vulnerably susceptible to SQL injection attacks [14] [33][47]. Before moving further, one must go through some fundamental definition for better understanding of the SQL injection attacks on web application and its underlying database [49].

**Vulnerability:** Vulnerabilities are the impuissance, loopholes, bugs or fault/imperfection in the subsisting system.

**Attack:** An attack is an illicit access i.e. a method to exploit vulnerabilities.

**Threat:** a series of events that utilizes the system in an unauthorized way compromising the principles of information security i.e. confidentiality, integrity and availability of the system.

**Risk:** Impact of the threat.

Albeit many researchers and practitioners have done the survey on SQL injection attacks against database but a detailed survey is done to elaborate the other aspects of attacks against database. In this paper an endeavour is done to provide the taxonomy of SQL Injection Attacks against database of a web application. This repository is the relegation scheme of attacks which includes- Research papers, white papers, technical reports and web sites. It can become a vital auxiliary in designing the security for web application and its underlying database.

The rest of paper is organized as follows: Section 2 describes the architecture of web application. Section 3 covers SQL injection attacks preview. Section 4 explains SQL injection classification and section 5 have explorative study of SQLIA. Findings of the study are described in Section 6. Section 7 briefs the security mechanisms. Finally section 8 concludes the paper.

## II. ARCHITECTURE OF WEB APPLICATION

For the better understanding of SQLIA one must have the cognizance of web application architecture. Web applications are a set of web pages and programs which reside on a web server [17]. The inputs provided by the user are sent to the server in the form of parameter string. These inputs are used to engender SQL query to retrieve information from the database. An authorized user can access it over the cyber world or over a public network and store the data in the database. A web application utilizes a web browser as an interface to extract the data from database server to accommodate the queries placed by the users [11][48]. Every web application is predicated on 3-tier architecture consisting of three layers [48]. Each layer can run potentially on a different machine and each layer should be independent of other layers. The three layers are

**Presentation Layer:** Presentation layer contains presentation logic. It is the top most level of application and handles the interactions with users. Its main function is to receive input from the user and provide the result in a convenient way that user can facilely understand.

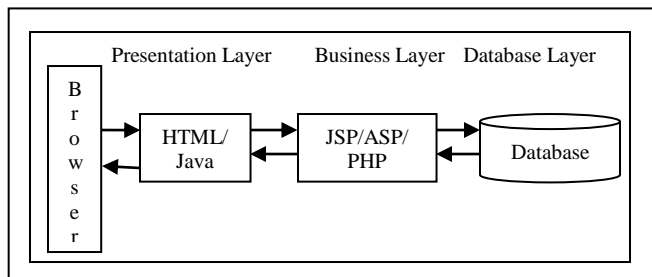


Fig. 1. Architecture of web application

**Business Layer:** This layer is present in between presentation layer and database layer. It is a logic layer which consists of a set of rules for processing the information between two layers. It contains application process commands which retrieves the data from database and sends to presentation layer for viewing the data. This tier can be programmed in any server scripting language like JSP, PHP, and ASP etc.

**Database Layer:** This is a physical storage layer for data persistence. It manages all access to database and file system. Information is stored and retrieved from database. It is then passed back to the presentation layer for processing and eventually back to the user. The main function of this layer is to provide access to authorized user and restrict the maleficent user.

**Working principle of architecture:** The presentation layer receives the request from web browser, processes it and then passes the dynamic part to the business layer which processes server scripting languages. All requests for database access are passed to the database server. The result is then dispatched to web browser as web pages. This architecture is easy to maintain and all the components are reusable. For the faster and smooth functioning, all the layers are governed and managed by different groups of experts. Web designer looks after the presentation layer. Software engineer does the logic and database administrator manages the database servers.

## III. SQL INJECTION ATTACK- PREVIEW

The focus on implementing the functionality rather than security from internal and external environment causes the susceptibilities in web applications. These are described as most solemn threat for web application as it may allow attacker to gain access to the web application and its underlying database. The potential of attacker perforates the system to extract sensitive information. Exploitation of loopholes in the design breaches the fundamental principles of information security i.e. confidentiality, Integrity, authenticity and availability of information [39][50]. Information stolen is loss of confidentiality. Loss of integrity takes place when data is modified in an unexpected manner. The Denial of service takes place when information is expunged for genuine user. Loss of authenticity means when information is accessed by unauthorized user. These attacks are application level attacks which are not obviated by firewalls [35][36][46]. It is hard to detect SQL injection prior to its impact. In most of the cases the unauthorized activity is performed through a valid user credentials for accessing the critical section of database of web application. The database servers are convinced that injected code is syntactically as valid as a SQL code. The inputs provided by the user form the dynamic SQL query to access the backend database. If these inputs are not opportunely sanitized, they can cause the web application to generate unintended outputs. The basic underlying fact is that SQL injection attacks are very easy to execute without any professional training.

Consider the following SQL statement.

```
SELECT Emp_info from Employee where E_name='abcd'
AND E_id = '12345';
```

The above query will provide the information about a particular desired employee for supplied input values but a SQL expression with injection will deport differently because the logic of the query is transmuted by the attackers, as

```
SELECT Emp_info from Employee where E_name='abcd'
AND E_id = '12345' OR '1'='1';
```

Because of an injection statement (OR '1'='1') the list of all the employees from the table in lieu of a desired output exposes the whole database. Such types of susceptibilities form the attacks [11]. Example expounded above is a very fundamental injection attack. The professional attacker uses very logical and resourceful keywords to extract the data from database servers.

## IV. ORGANIZATION OF SQL INJECTION ATTACKS

The objective of SQL Injection Attack (SQLIA) is to penetrate the database system into running inimical code that can reveal confidential information. This is done by injecting the SQL queries and expressions as an input string to gain an unauthorized access. SQL injection is a threat that leads to a high level of compromise - conventionally the ability to run any database query. It is web-predicated application level attacks that connects to backend database and bypass the firewall. The advantage of insecure code and deplorable input validation is clinched by the attacker to execute unauthorized SQL commands. On the substratum of attacks against the



database management system, SQL injection attack can be classified as [7][10][16][43][45][46].

SQL Injection attacks against web application databases can be divided in four sections as follows:

- 1) Code Injection
- 2) SQL Manipulation
- 3) Function Call Injection
- 4) Buffer Overflows

Code injection is when an attacker inserts new database commands or SQL statements into user's statement. The manipulation involves modifying the SQL statement through set operations or altering the WHERE clause to return a different result. When attacker injects a customized function or already existing function into a SQL statement, that type of injection is known as function call injection. These are used to manipulate data in the database. Buffer overflows is a subset of function call injection. In several commercial applications vulnerabilities exist in database functions that may result in a buffer overflow.

#### a) Code Injection

In code injection, attacker attempt to add additional SQL statement or commands to the existing SQL statements.

##### Original query

```
SELECT Emp_salary FROM Employee WHERE E_name='kushagra' and E_id='abc123';
```

##### Output

A salary statement of desired employee is extracted from the database.

##### Injection query

```
SELECT Emp_salary FROM Employee WHERE E_name='kushagra' and E_id='abc123'; DELETE FROM Employee WHERE username = 'kushagra';
```

The above query uses the stored procedure command which is inserted between original query and injected query so that the both queries executes in a single run as a single query.

##### Output

Information about desired employee is deleted from the database due to additional command.

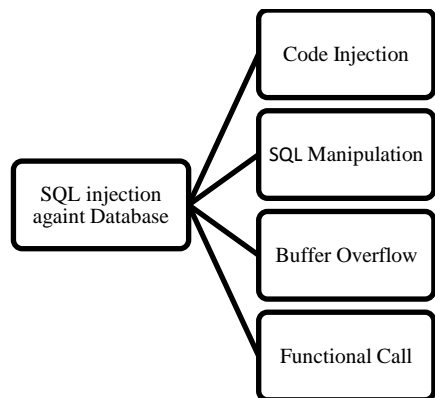


Fig. 2. Classification of SQLIA against database

#### b) SQL Manipulation

A very familiar kind of SQL Injection attack is SQL manipulation. The most significant examples of this type is by adding elements to the WHERE clause with set operators like UNION, INTERSECT etc. Or comparators like OR, <, > and many more. The simplest example is login authentication that a web application may check.

##### Original query

```
SELECT * FROM Employee WHERE username='kushagra' and Password='abc123';
```

##### Output

A desired employee details are returned by the database.

##### Manipulation query

```
SELECT * FROM Employee WHERE username='kushagra' OR '2' > '1'and Password='-----';
```

Based on operator precedence, the clause WHERE is true for every entry and (-----) regarded as comment consequently ignored by server granting access to attacker.

##### Output

Information about all users is received without authorization.

It can also extract information about all the users using union query.

```
SELECT * FROM Employee WHERE username like '%kushagra'; UNION SELECT username FROM users WHERE username like '%';
```

#### c) Function call injection

Function call injection is the insertion of functions which executes with a SQL statement. Functions which are marked as "PRAGMA TRANSACTION" are executed as part of a SQL SELECT statement. Using INSERT, UPDATE, or DELETE in SQL statement attacker is able to modify data in the database. By using custom functions an attacker can send information to a remote computer or execute other attacks on the database server.

*Example:* A custom application has the function NEWUSER in the custom package MYAPP. The developer marked the function as "PRAGMA TRANSACTION", since it is marked "PRAGMA TRANSACTION", it can write to the database even in a SELECT statement

```
SELECT TRANSLATE (' // myapp.newuser ('admin', 'newpass') // ', '12345ABCDE', 'abcde56789') FROM dual;
```

This type of injection can even exploit the simplest SQL statement at runtime.

#### d) Buffer Overflow

When the process of writing data to buffer overruns the boundary of buffer and overwrites adjacent memory then an abnormality is recorded known as Buffer overflow. These injection attacks are triggered either by variations in inputs or by amending the method of program operation. The standard database functions which can be exploited through a SQL

injection are `tz_offset`, `to_timestamp_tz`, and `bfilename`. A buffer overflow attack is executed using the function injection methods. The very effective buffer overflow attack is denial of service [44][52].

*Example:* The web process gets hanged until connection to the client is terminated.

## V. EXPLORATIVE STUDY OF SQLIA

In this section the above briefed attacks are discussed in detailed with their code injecting mechanism, types and subtypes with appropriate SQL statements and queries.

### A. Code Injection

*Injection Mechanism:* The attacker attempts to manipulate the SQL statement by injecting additional code, operator sub query in to original query.

The code injection attacks can be further divided in to four types.

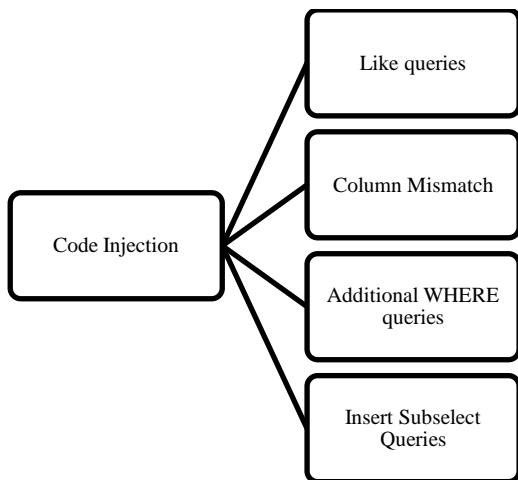


Fig. 3. Types of code Injection attacks

*a) Like queries:* Like query is used to compare a value to similar values using wildcard operators. Basically two wildcards are used

- i) % represents zero, one or multiple characters
- ii) underscore ( \_ ) represents a single number or character.

The attacker manipulate the SQL statement using these wildcards. A Denial of Service attack can be launched with a few changes in a LIKE query by overloading the database [7][44][51].

*Example:*

```
SELECT Employment_No FROM Employee WHERE E_Name LIKE 'A%';
```

It provides the list of all employees starting with name A or having name an alphabet A

*b) Column Mismatch:* This particular attack occurs when there are errors like mismatch operand type or “Queries containing a UNION operator must have same number of

expressions”. To gather the information of this type of injection, consider a random SQL query.

*Example:*

```
SELECT product_name FROM all_products WHERE product_name like '&Chairs&'
```

It provides the name of product according to the query input but attacker changes the logic of the query it becomes

```
SELECT product_name FROM all_products WHERE product_name like '' UNION SELECT ALL FROM Objects WHERE '' = '';
```

Above query would give errors that indicate that there is mismatch in the number of columns and their data type in the union of Objects table and the columns that are specified using ALL. The error is caused by the injected string. Another error is because the number of columns is not matching. This information is enough to penetrate the database of any web application [40][43].

*c) Additional where queries:* The attacker can also extract the information from the message displayed by the database while using an additional WHERE clause with the input string.

*Example:*

```
SELECT FName, LName from Employees WHERE City= 'Delhi' AND Country='India';
```

This query will provide first name and last of all the employees working in a particular city & country. The modified query with injected additional WHERE clause will be

```
SELECT FName, LName from Employees WHERE City= 'Nosuchcity' UNION ALL SELECT SomeField from SomeTable WHERE 1 = 1 AND Country = 'USA';
```

Due to this insertion clause the detailed error message displayed by the database reveals the name of table and column name in error message- like Invalid column Name 'Country' because 'Table1' does not have a column called 'Country'. It exposes the database with table name

*d) Insert subselect Queries:*

The insertion of a sub query into a query can also help the attacker to access all the records of the database

*Example:*

```
SELECT E_Name FROM Employee WHERE Employee Employment_No = (SELECT work_in.Employment_no FROM work_in WHERE D_No= 23)
```

The result of above query will provide the result to find the name of all employees working in department no 23. If the query is executed then the result is displayed otherwise there will be an error message like “subselect returned too many rows”. Attacker can go through all the record using NOT operator. This attack is possible where users are allowed to edit user information [7][52].

## B. SQL Manipulation

*Injection Mechanism:* The attacker attempts to manipulate the SQL statement by injecting the code in one or more conditional statements.

These SQL manipulation attacks have four types as shown in the figure given below.

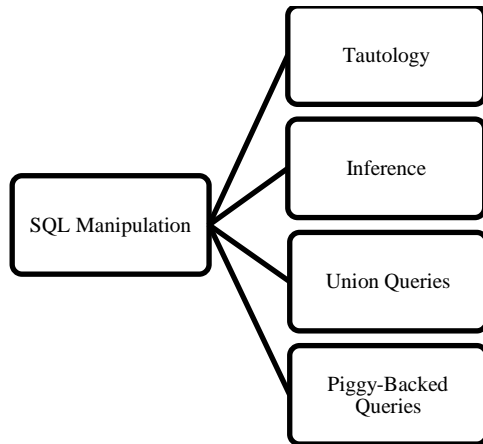


Fig. 4. Types of SQL manipulation

*a) Tautology:* In this type of SQLIA, an attacker exploits an injectable field that is used in a query. The query always returns result upon evaluation of a WHERE conditional parameter. The aim of this attack is to inject malicious codes into one or more conditional statements which are always evaluated to be true. All the rows in the database table targeted by the injected query returns the conditional WHERE into a tautology.

*Example:* It allows an attacker to log on to application without supplying a valid user name. The attacker submits

'Abcd' OR '1'='1'/'A'='A' in login field and '-----' in password or pin field.

The resulting query is:

```
SELECT Emp_id FROM Department WHERE  
Login='Abcd' OR 'A'='A' AND pass='-----';
```

As the condition (OR A=A) is always true and ---- is used for comments, it converts the entire WHERE clause into a tautology [38][40][45].

*b) Inference:* when the attacker tries extracting information from not enough secured backend database via error messages then inference injection is executed. A different method must be used by the attacker to get the response from the database since database error messages are not available without executing a query or statement. The error messages displayed by database may become useful tools for attacker to plan an attack. In this situation, the attacker injects commands and then observes the change in function/response of application. By carefully observing

the behaviour of application vulnerable parameters can be extrapolated by attacker with added information about the database.

*Example:* SELECT Employee from Bankers where E\_number=' \$%^&\*!@# and E\_id='AZ+=79%':

The input provided in query is incorrect & results in an error message. The displayed error message should be like

"Microsoft OLEDB provider for SQL Server (0x80040E07) error converting nvarchar value 'E\_number' to column of data type int".

The Information about version and schema of backend database and is revealed to attacker to plan further Inference attacks [10][40][43].

*c) Basic union queries:* This type of attack is also called statement injection attack. The attacker tricks the database server to return data that which is not intended by the authentic user. The vulnerable parameters are exploited with the help of keyword UNION, which is used to join original query and an injected query. The attacker controls the injected query completely, to retrieve information from database. The output of this attack causes the database return values which is union of two queries

*Example:* SELECT Basic\_info from Employee where user E\_name='xyz' and E\_id='-----'; UNION SELECT

Salary\_info from Employee where Emp\_id='1234';

The first part of the query gives null values but second part of the query returns the information of employee having id 1234 [10] [16][40][51].

*d) piggy- backed queries:* In this category the attacker's aim is not to modify the query. The addition of distinct queries with the valid query is desirable. When database receives multiple queries, shows extremely harmful results like deletion or removal of information with a harming intention.

*Example:* SELECT Basic\_info from Employee where user E\_name='xyz' and E\_id='1234'; DROP table Employee;

Here two queries are separated by delimiter (;). So both the queries get executed. After execution of first query the database proceeds for the injected second query. When second query executes, it will drop table 'Employee', from database which results in potential damage to important information. In the similar manner there are various other types of queries like inserting new employees to table etc. [10][40][45].

## C. Function Call Injections

*Injection Mechanism:* The attacker attempts to manipulate the SQL statement by inserting database functions or custom functions into a vulnerable SQL statement [43][53]. This type of injection attack has two types role function and system stored procedure

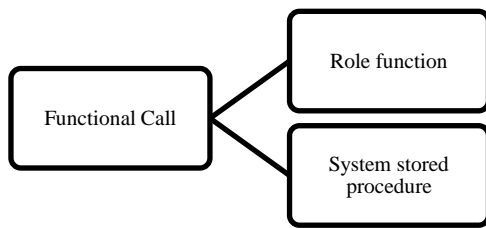


Fig. 5. Types of function call injection

the knowledge of this type of injection, Consider the Webportal of a company which provides press release on regular intervals.

*Example:*

`www.company.com/PressRelease.php?pReleaseID`

The corresponding SQL statement used by the application would look as under, if pRelease is 7

```
Select title, description, releasedate, body from  
pressRelease WHERE pRelease ID=7
```

All the information requested corresponding to the 7th press release is returned by the database server. This information is in a HTML page (understood by the browser) and provided to the user. Certain changes in parameters in the address of a HTML page by attacker can change the role of a query. Like

```
Select title, description, releasedate, body from pRelease  
WHERE pRelease ID=7 AND 2 > 1;
```

```
Select title, description, releasedate, body from  
pressRelease WHERE pRelease ID=6;
```

If the application still returns some document i.e. it is susceptible to SQL injection attack and attacker can plan the attack accordingly [7][47].

*a) Stored system procedures:* In this type the in-built stored procedures are attacked by attacker using malicious sql injection codes. Every database uses the stored procedures. The knowledge of running backend server allows the attacker to penetrate the system using the stored procedures. This attack is like Piggy Backed queries attack in which stored procedure command is inserted inbetween original query. The both queries execute as one.

*Example:* `SELECT Emp_ Salary from Employee where username='abcd'; SHUTDOWN; and password='12345';`

In above query SHUTDOWN is a stored procedure which causes shutdown the database. The admittance to system store procedures depends on the access privileges to user by the application [10][40][51].

#### D. Buffer Overflow

*Injection Mechanism:* The attacker attempts to manipulate the SQL statement by using database functions which are susceptible to buffer overflows. In several databases some database functions are vulnerable to buffer overflows that can be exploited through a SQL injection attack. A very efficacious denial of service attack hangs the process until the

connection is terminated with the above said attack [44][52][53].

During the preparation of this repository some other methods are encountered. These methods do not fit under any attack schema but can be used to inject malevolent code in to user's code for executing SQL injection attacks. The methods are explained as under [17][35][36][46][48][52].

*a) Sophisticated Matches:* One of the prevalent signatures utilized by such mechanisms is some remotely variant on the famous OR '1'='1' attacks. Sophisticated matches technique uses alternative expression of OR '1'='1'. For example: OR 'Unusual' = 'Unusual', OR 'Simple' = 'Sim'+ 'ple', OR 2 > 1 and OR 'Simple' BETWEEN 'R' AND 'T' all have the same effect as OR '1'='1'.

*b) Hex Encoding:* Hex encoding technique uses hexadecimal encoding to represent a string. For example, the string 'SELECT' can be represented by the hexadecimal number 0x73656c656374, which most likely will not be detected by a signature protection mechanism.

*c) Char Encoding:* Char encoding technique uses build-in CHAR function to represent a character. For example, the string 'SELECT' can be represented by the CHAR function as char (73)+char (65) +'LECT', which make it very arduous for detection system to build a signature that match it.

*d) In-line Comment:* In-line comment technique obscures input strings by inserting in-line comments between SQL keywords. For instance, `/**/UNION/**/SELECT/**/name designation` can elude detection from signatures that expects white space between SQL keywords.

*e) Dropping White Space:* Dropping white space technique obscures input strings by dropping white space between SQL5 keyword and string or number literals. For example, OR 'Simple'='Simple' works precisely the same way as OR 'Simple' = ' Simple', but has no spaces in it, make it capable of eschewing any spaces predicated signature.

*f) Break Words in the Middle:* With MySQL, the in-line comments would not work as supersession for a space. The in-line comments can be utilized in MySQL to break words in the middle, for Instance: `UN/**/ION/**/ SE/**/LECT/**/` is evaluated as UNION SELECT.

## VI. FINDINGS OF STUDY

The study of the SQL injection attacks (SQLIA) against database centric web application concludes that these attacks possess the great threat. Unfiltered user inputs invite these attacks. Bypass the authentication process sanctions the attacker to postulate all privileges associated with genuine user. The Retrieval of personal and sensitive information is highly desirable by the attackers. Mostly SQL injection attacks are done to steal the sensitive information. By knowing the version and type of the database utilized by the user makes it facile for the attacker to craft a query. Sometimes the goal of attacker is to integrate the information in a database to mark one's identity; if it is done for enjoyment then no harm is caused but an attacker with destructive intention can delete the whole database. By remote command execution attacker can even

shutdown the database. Based on above discussion some findings are listed as under

A. *Most common reasons are behind SQLIA*

- a) *Mismatch Data type*
- b) *Accounts with more access*
- c) *Insufficient input validation*
- d) *Detailed Error messages*
- e) *No sanitization of data sent to the server through URL.*

B. *Privileges gained on successful SQLIA*

- a) *Access to Database schema.*
- b) *Disclosure threats of the asset.*
- c) *Gain access to host and internal network.*
- d) *Exploitation of susceptibilities of the web application.*
- e) *Privilege escalation.*
- f) *Impose deception and usurpation threats.*

C. *Methods for eschewing SQLIA*

- a) *Avoid building dynamic SQL expression from user input.*
- b) *Length of input string must be constrained.*
- c) *Avoid using query delimiter, SQL keyword, character data string delimiter and single line comment in user input.*
- d) *Use different Database account for different calibers of privileges.*
- e) *Error messages must be customized to hide the details of injectable parameters.*
- f) *Use parameterized queries for Database access.*
- g) *Use stored procedures to avoid direct access of Database.*
- h) *Evade building SQL statements from cookie and HTTP variables.*

## VII. SECURITY MECHANISMS

When good programming habits and eschewing methods are not sufficient for the avoidance of SQL injection attacks then some manual and automated security mechanisms are applied for the protection of web application database. These are very inimical attacks which target the most valuable assets of web application. The consequences of SQL injection attacks range from modification in data to denial of data. Researchers have proposed several techniques to contravene SQL-injection attacks which include - Code review, Defensive programming, Software hardening techniques, and Hardware extensions feature in modern processors, Attack detection and containment mechanisms and many more. These approaches are not sufficient to counter the problem of SQL injection attacks. Some security mechanisms are suggested to faceoff SQL injection attacks.

A. *Detection and Prevention tool*

When the techniques like defensive coding & operating system hardening [13] are not enough to stop SQLIA, some

tools are required for detection and prevention of web application and its underlying database. Prevention designates to evade unauthorized user (attacker) from accessing any component of system or data. Prevention tools [26][30][34] are runtime analysis for checking susceptibilities by placing a validation checker between web server and database server. It additionally averts the attacks that capitalize on type mismatch, sanitization of inputs and input sources withal [20][22]. Detection determines whether someone has attempted to break into your system, if yes, then up to what extent of damage may have been done. Detection tools are implemented in two approaches [18][31][37] – Static technique and Dynamic technique. Static technique [33] is applied directly without running the code, it includes approaches [6][27][28] like – Pattern Matching, Lexical Analysis and Parsing (includes type qualifier, dataflow analysis, taint analysis and model checking). Whereas dynamic techniques [33] include approaches like- Fault injection, Fuzzy testing, Dynamic taint & Sanitization of inputs [5]. Most of the approaches [31] are not implemented yet as a tool. So the scope of developing prevention tool, detection tool or the combination of both prevention and detection tool is on the cards for the researchers.

B. *Instruction Set Randomization*

A technique to counter SQL- injection attack, is instruction set randomization (ISR) [1][4][12][46]. The fundamental idea behind this approach is that attacker doesn't know the language spoken by the runtime environment on which an application runs, so a SQL-injection attack will ultimately fail because the foreign code, however injected, is written in a different language. An ISR to SQL injection is straightforward approach. It randomizes both the underlying runtime environment the SQL parser and the SQL program (the template that the Web application uses). A simple approach for randomizing the SQL grammar consists of appending a random numeric tag (the randomization key) to each statement and operator in SQL .This can be an efficacious way of ceasing injection attacks, but it typically requires extensive modifications to the runtime environment.

C. *Intrusion Detection System and Proxy Server*

Most of the SQL injection attacks are application level attacks. A model can be built to filter the SQL queries which work as security layer between database server and web server to filter the queries at run time[9][12][19][24]. Similarly a Proxy filter can be developed to intercept the HTTP request and enforce input validation so that malevolent SQL expression would be averted to send to database server [15][23][29]. Many models have been proposed by researches based on different approaches like machine learning, intended structure of SQL verbalizations, data flow analysis, identification of critical points and many more but are not sufficient to avert SQL injection attacks [29][31]. So an incipient hybrid model can be proposed which may be the combination of several approaches to counter injection attacks.

D. *Threat Model*

Threat modeling [2][3][8][13] is a procedure for optimizing Security by identifying objective, susceptibilities and then defining countermeasure to avert the effects of threat to the

system [19].The threat modeling process customarily involves identifying information sources to be bulwarked, ingress points or access points to the system's assets, analyzing the threats, evaluating the associated risks and developing mitigation strategies[25].There are three different approaches of threat modeling – Asset centric (Find risks associated and rank the risks), Attacker centric (Find level of harm, Evaluating damage potential and risk rating) and Software centric (Decomposes the application to identify the threats and then mitigate the threats).Many more threat models [41][42] or a hybrid model (amalgamation of two or more models) can be proposed to filter the susceptibilities and malevolent SQL verbal expressions from different input sources and mitigate the attacks.

### VIII. CONCLUSION

In this paper the detailed analysis is presented on various types of SQL injection attacks with related vulnerabilities and injection mechanisms. SQL Injection is a common technique that attackers employ on web based data centric applications. These attacks modify the SQL queries in a manner to alter the behaviour of application. This paper also provides the taxonomy of mechanisms for avoidance, prevention and detection from these attacks. In future work, a Threat model will be proposed as a security mechanism for securing database of web applications.

### REFERENCES

- [1] E.G. Barrantes,"Randomized Instruction Set Emulation to Disrupt Binary Code Injection Attacks," Proc. 10th ACM International Conference on Computer and Communication Security, pp. 281–289, 2003.
- [2] F.Swiderski and W.Snyder,"Threat modeling" (Microsoft press, 2004)
- [3] Threat risk modeling,"A Guide to Building Secure Web Applications and Web Services," (2.0 Black Hat edition, p. 38-51,July 2005)
- [4] A. Sovarel, D. Evans and N. Paul,"Where's the FEEB? :The Effectiveness of Instruction Set Randomization," Proc. Usenix Security Symposium, Usenix Association,pp. 145–160, 2005.
- [5] W.G.Halfond and A.Orso.Amnesia,"Analysis and monitoring for neutralizing sql injection attacks," Proc.20th ACM Int. Conf. on Automated Software Engineering., Long Beach, California,USA, p.174,2005.
- [6] A McClure and Ingolf H.Kruger,"SQL DOM: Compile TimeChecking of Dynamic SQL Statements," Proc. International ACM Conference of Software Engineering, pp. 88-96, May 2005.
- [7] Sagar Joshi,"SQL injection attack and defense: Web Application and SQL injection," 2005.
- [8] P.Torr,"Demystifying the threat modeling process"IEEE transaction on Security & Privacy, 3(5), pp. 66- 70, 2005.
- [9] F. Valeur, D. Mutz and G. Vigna,"A Learning-Based Approach to the Detection of SQL Attacks" Detection of Intrus Malware and Vulnerability Assessment Proceedings, Volume: 3548, pp.123-140,2005.
- [10] W.G.Halfond,J. Viegas and A. Orso,"A Classification of SQL-Injection Attacks and Countermeasures," Proc. IEEE Int. Symposium on Secure Software Engineering, Washington,DC,USA, March 2006.
- [11] Z.Su and G. Wassermann,"The Essence of Command Injection Attacks in Web Applications," Proc. 33rd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages,Charleston, South Carolina, USA, pp. 372-382,2006.
- [12] Y.Weiss and E.G. Barrantes,"Known/Chosen Key Attacks against Software Instruction Set Randomization," Proc. Annual Computer Security Applications Conf. (ACSAC),pp. 349–360,2006.
- [13] E. A.Oladimeji and S.Supakkul, L. Chung,"Security threat modeling and analysis: a goal-oriented approach,"Proc. of the 10th IASTED International Conference on Software Engineering and Applications (SEA), pp.13-15,2006.
- [14] Jonse Fonseca,"Testing and comparing web vulnerability scanning tool for SQL and XSS attacks," Proc. 13th IEEE symposium on Pacific Rim Dependable Computing,pp 365-372,2007.
- [15] Xiang Fu and Xin Lu,"A Static Analysis Framework For Detecting SQL Injection Vulnerabilities," Proc. 31st IEEE Annual International Computer Software and Application Conference, pp. 87-96,24-27 July 2007.
- [16] San-Tsai Sun, Ting Han Wei, Stephen Liu, and Sheung Lau,"Classification of SQL Injection Attacks,"Electrical and Computer Engineering, University of British Columbia, November 2007.
- [17] D.Stttard and M.Pinto,"The Web Application Hacker's Handbook:Discovering andExploiting Security Flaws,"(Wiley publication, 2007)
- [18] M.Cova,D.Balzarotti,V.Felmetsger and G.Vigna, "Swaddler: An Approach for the Anomaly-based Detection of State Violations in Web Applications", the Recent Advances in Intrusion Detection (RAID), Gold Coast, Australia, pp.63 – 86, 2007.
- [19] K.Kemalis and T.Tzouramanis,"SQL-IDS:A Specification based Approach for SQL Injection Detection," Symposium on Applied Computing. ACM, pp.2153-2158, 2008.
- [20] Monticelli, F,"SQL Prevent," University of British Columbia (UBC) Vancouver, Canada.2008.
- [21] Symantec,"Symantec Report on Underground Economy," pp.9-12,Symantec, 2008.
- [22] Jin-Cherng Lin, Jan-Min Chen and Cheng-Hsiung Liu,"An Automatic Mechanism For Sanitizing Malicious Injection," Proc. 9th IEEE International Conference For Young Computer Scientists,pp.1470-1475,18-21Nov 2008.
- [23] Anyi liu and yi yuan,"SQLProb: A Proxy based Architecture towards preventing SQL injection attacks,"ACM, pp.2054-2061, March 2009.
- [24] Anglos D. Keromytis,"Randomized Instruction sets and run time Environment,"IEEE Transaction on Security & Privacy, pp.18-25, Jan/Feb 2009.
- [25] F.Swiderski and W.Snyder,"Threat modeling" O'Reilly Media, Inc., 2009.
- [26] P.Bisht, P.Madhusudan and V.N. Venkatakrishan," 'CANDID: Dynamic Candidate Evaluations for Automatic Prevention of SQL Injection Attacks," ACM Transaction on Information System Security, pp.1-39,2010.
- [27] Ivano Alessandro Elia, Jose Fonseca and Macro Vieira,"Comparing SQL Injection Detection Tools using Attacks Injection: An Experimental Study,"Proc. 21st International Symposium on Software Reliability Engineering,pp.289-298,1-4 Nov 2010.
- [28] Jeom-Goo Kim,"Injection Attack Detection using the Removal of SQL Query Attribute Values,"Proc. International Conference in Information Science and Application(ICISA),pp.1-7,26-29 April 2011.
- [29] Chai Wenguang, Tan Chunhui and Duan Yuting,"Research of Intelligent Intrusion Detection System Based On Web Data Mining Technology," Proc. IEEE 4<sup>th</sup> International Conf. on Business Intelligence and Financial Engineering,pp.14-17,17-18 Oct 2011.
- [30] Indrani Balasundram and E .Ramaraj,"An Authentication scheme for Preventing SQL Injection Attack Using Hybrid Encryption," (PSQLI-HBE),53(3), pp.359-36, 2011.
- [31] Atefeh Tajpour, Suhaimi Ibrahim and Mohammad Sharifi,"Web Application Security by SQL Injection Detection Tools," International Journal of Computer Science, 9(2), pp.332-338,2012.
- [32] Avanish Kumar Singh and Sangita Roy,"A network based Vulnerability scanner for Detecting SQLI Attacks in Web Applications," Proc. 1st Int. Conf. on Recent Advances in Information Technology (RAIT-2012), Dhanbad, India, pp.585-590,15-16 March 2012
- [33] Sruthy Mamadhan, Manesh T and Varghese Paul," SQLStor: Blockage of Stored Procedure SQL Injection Attack Using Dynamic Query Structure Validation," Proc. IEEE 12<sup>th</sup> International Conf.on Intelligent Systems Design and Applications(ISDA),pp.240-245,27-29 Nov 2012.



[34] Debabrata Kar and Suvasini Panigrahi, "Prevention of SQL Injection Attack Using Query Transformation and Hashing," Proc. IEEE 3rd International Conf. on Advance Computing, pp.1317-1323,22-23 Feb 2013.

[35] Chad Dougherty, "Practical Identification of SQL Injection Vulnerabilities," Produced for US-CERT© 2012

[36] Cenzic vulnerability report 2013.

[37] Jaskanwal Minhas and Raman Kumar, "Blocking of SQL Injection Attacks by Comparing Static and Dynamic Queries," International Journal Computer Network and Information Security, vol.2, pp.1-9, 2013.

[38] Chandershekhar Sharma and S.C. Jain, "SQL Injection Attacks on Web Application," International Journal of Advanced Research in Computer Science and Software Engineering, 4(3), pp.1268-1272, 2014.

[39] F. S. Labs, "Threat report" Last Accessed: 27-3-2014.

[40] Chandershekhar Sharma and S.C. Jain, "Analysis and Classification of SQL Injection Vulnerabilities and Attacks on Web Applications," Proc. IEEE Int. Conf. on Advances in Engineering & Technology research (ICAETR-2014), Dr. virendra Swarup group of institutions, Unnao, India, pp.1-6, August 2014.

[41] Adam Shostack, "Threat Modeling: designing for security" Wiley publication, 2014.

[42] Satapathy Soumya Ranjan, "Threat Modeling in Web Applications," Thesis-NIT, Rourkela, 2014.

[43] C. anely, "Advanced SQL injection in SQL server applications" White paper.

[44] Kevin J.Houle, "Trends in Denial of Service Attack Technology," Whitepaper.

[45] S.Mcdonald, "SQL Injection: Modes of Attack, Defense, and Why It Matters," White paper.

[46] [www.salientsecurity.com](http://www.salientsecurity.com)

[47] [www.securitydocs.com/librarys](http://www.securitydocs.com/librarys)

[48] [www.owasp.org/index.php/Top\\_10\\_2013](http://www.owasp.org/index.php/Top_10_2013).

[49] [www.bcs.org/upload/pdf/infosec-mgt-principles.pdf](http://www.bcs.org/upload/pdf/infosec-mgt-principles.pdf)

[50] [www.GovernmentSecurity.org](http://www.GovernmentSecurity.org).

[51] [www.spidynamics.com/paper/SQLIWhitePaper](http://www.spidynamics.com/paper/SQLIWhitePaper)

[52] M.Howard and D.LeBlanc, "Writing Secure Code," Microsoft Press, Redmond, Washington.

[53] S.Kost, "An introduction to SQL injection Attacks for Oracle Developers.pptx

#### AUTHOR'S PROFILE



performance Computing and Threat modeling



Dr S.C.Jain has done his PG in Computer Science and Technology from IIT Roorkee and PhD in VLSI design from IIT, Delhi. He has served Defense Research and Development organization, Bangalore, India and presently working as Professor, Computer Science and Engineering, Rajasthan Technical University, Kota, India. His Research interests are VLSI design, Real Time Embedded System, Reversible Computing and High Performance Computing.



Dr. Arvind K Sharma holds PhD degree in Computer Science. He has more than 13 years of work experience in academics field. He has published more than 27 Papers in various National, International Journals and Conferences. He has authored & co-authored almost 5 books. He has received Best Paper Award in International Conference, 2012 held in Thailand. He has visited Thailand & Dubai for attending International Conferences. He has participated as Speaker & Keynote Speaker in many National and International Conferences. He is a Sr. Member of numerous academic and professional bodies i.e. IEEE, WASET, IEDRC, IAENG Hong Kong, IACSIT Singapore, UACEE UK, ACM, New York. He is a Member of Technical Advisory Committee of many International Conferences in India and abroad. He is also Editorial Board Member & Reviewer of several National and International Journals. Besides it, he is serving as a recognized Research Supervisor to guide Masters and PhD scholars in many Universities of Rajasthan, India. His area of interest includes Web Usage Mining, Web Intelligence Applications, Web Data Mining, Big Data Analytics and Machine Learning Tools.

# Corrupted MP4 Carving Using MP4-Karver

Ahmed Nur Elmi Abdi, Kamaruddin Malik Mohamad, Yusooof Mohammed Hasheem, Rashid Naseem, Jamaluddin, Muhammad Aamir

Faculty of Computer Science and Information Technology  
Universiti Tun Hussein Onn Malaysia  
Batu Pahat, Malaysia

**Abstract**—In the digital forensic, recovery of deleted and damaged video files play an important role in searching for the evidences. In this paper, MP4-Karver tool is proposed to recover and repair the corrupted videos. Moreover, MP4-Karver extracts frames from video for automatically screen-video to detect illegal cases instead of targeting or watching complete video. Therefore, many existing approaches such as Scalpel's method, Garfienkel, Bi-Fragment Gap Carving, Smart Carving and Frame Based Recovery attempts to recover the videos in different ways, but most of the recovered videos are usually not complete playable. The proposed MP4-Karver focuses on recovery of video files and repair corrupted videos to complete and playable. Experimental results show that the proposed MP4-Karver effectively restores corrupted or damaged video for an improved percentage of the video restoration compared with existing tools.

**Keywords**—Digital forensic; File carving; Repairing; Corrupted; Frame extraction

## I. INTRODUCTION

Recovery of corrupted or damaged video files, is playing a crucial part in digital forensics for detecting illegal cases. In the last decade, for video recording the digital devices with built-in cameras like surveillance camera, CCTVs and mobile cameras are increased enormously [1]. Similarly, the large amount of recorded video data has been made due to the common usage of these digital devices. Although digital devices have their own internal memory or external devices like external hard disk, flash memory and multimedia card.

Besides the increase digital devices and data storage, the numbers of users of these devices are also increased, according to Pew Research Center Globalization review in which 26 out of the 36 countries surveyed [2]. The video data files stored in the digital storages can be easily fabricated and modified, thus the role of digital forensics has increased to collect, analyze and process the digital evidences.

To recover the video data from digital devices, File carving is a powerful method used in digital forensic by taking advantage of the video file's structure. Studies on the file carving have targeted the restoration of different file types like pdf, zip, and image files[3]. But there are small number of methods on carving of the file type like video and audio files [4]. These techniques uses header and footer to recover the video files[5]. However, when video file structure is not present or corrupted (fragmented), then by using the existing techniques the recovering possibility of that file is below than 50% [6].

After retrieving video data, each video file is to be screened manually in order to find the presence of any illegal cases. The screening of each individual video a very stressful process and also a time consuming job, because a video file is a series of several still images called frames [7]. A few minutes of video might consist of thousands of video frames.

In this paper, MP4-Karver tool is proposed to recover and repair the corrupted MP4 videos and make it playable. Even if the MP4 video is produced from CCTV or any other device. The proposed tool performs better than the existing tools and techniques by having three main features. 1) Retrieving video. 2) Repairing corrupted video. 3) Frames extraction. All other tools have the ability to retrieve video but they are unable to repair the corrupted one. This tool can extract frames from video into image files, for analyzing video. The rest of the paper is organized as follows; Section 2 discuss the related work consists of an overview of digital forensic, file carving and existing method video recovery while Section 3 defines MP4 restoration and repairing framework. Section 4 presents experimental and result discussion. Finally, Section 6 concludes this paper.

## II. RELATED WORK

Digital forensics is an improving method and commonly refers to the process of investigating digital files. For digital forensic, file carving is powerful method to recover data from unallocated space and it depends upon the file structure instead of file's metadata [8]. Following are the different video recovery techniques.

The most common techniques are Scalpel's method, Garfienkel method, Bi-fragment gap carving technique, smart carving and Frame based recovery. Scalpel technique proposed a popular tool that does not depend on file system or metadata of the file for recovering a video file[9]. In this technique, it is necessary to have an indexing step to find the file header and footer from a whole disk image. Similarly, a restoration step is required to restore indexed header and footer. This technique can be used to restore non-fragmented files and cannot recover partially overwritten video files.

Another technique that is used to recover video is Garfienkel method [10]. This technique can restore video files even if the header or footer is fragmented. An extension of the signature based file restoration is used to recover damaged file. Like Scalpel technique, this technique does not also use the metadata of the file system.

In this technique, additional file size and length information is added to the header of the video. Therefore, based on the length information, a video file can be restored even when fragmentation happens and the file footer is removed. However, this technique fails when the allocated memory to the file that is to be recover, is overwritten.

Bi-fragment Gap Carving [11] is used to restore the video files. In this technique, a combination of region containing file header and footer are found and analyses. The difference between header and footer regions are measured and tested. If the differences do not qualify the default validation procedure, it is required to repeat the process until the gap passes the validation check. Nevertheless, this technique can only be used in a video file with two fragments and the main limitation is when the gap between the two file fragments is large.

Smart carving [12] is a technique proposed to restore files without considering the number of fragments in the file. It identifies the fragmented parts of the file, and then it performs permutation and combination process on that file. In this process the frames order of video identified and then restored it back to the original video file to check whether the recovered video is meaningfully and playable or not. However, there is a drawback of this technique. When a part of allocated memory to the video is overwritten, smart carving fails to retrieve that video.

A new technique known as Frame based recovery of damaged video file using video codec specification is recently proposed [13]. In this method a video file is separated into frames where every frame has at least a meaningful unit of a video. Each unit of video frame is encoded by using a codec description. Similarly, each frame has decoding header information at the start and end of a frame. Therefore, in this technique video carving is done by using video frames and decoding header information. However, the frame based recovery method can repair frames of the non-overwritten video files and partial overwritten video files [14]. Moreover, this technique cannot recover completely over-written videos.

In this paper, MP4-Karver tool is proposed to carving video files and repair corrupted videos. Then extract frames from videos for digital forensic analyzing.

### III. MP4-KARVER VIDEO METHOD

This section describes about the benchmark data sets and compared tools which are used to evaluate the performance of the proposed MP4-Karver against other video carving tools. Also, this section explains the methodology of the proposed tool for recovering and repairing MP4 videos.

#### A. Data sets

The standard data sets are used to evaluate in a digital forensic investigation are commonly bitwise-copies of full hard drives. Table 1 presents the statistics of three data sets that are used in the proposed tool.

- Digital Forensic Research Workshop challenge data set (DFRWS-2007-challenge.img).
- Baseline Carving Data Set (BCDS.raw).

- Level\_3-(Video.dd) from computer forensic tool testing (CFT)

TABLE I. STATISTICS OF THE DATA SETS

Data sets	Size	Number of files
DFRWS.img (BFRWS)	330 MB	8 MP4
BCDS.raw (BCDS)	237 MB	3 MP4
Level_3- Video.dd (Level)	48.5 MB	1 MP4

#### B. MP4 Video carving tool

There is a wide selection of video carving tools available ranging from expensive proprietary forensic tools (blade forensic) to open source tool (PhotoRec). A total of 3 file carving tools are selected for testing the proposed tool and they are listed in Table 2. The basis for tool selection criteria included: MP4 video file support, advanced carving features and tool availability. Each tool listed has the associated license, tool version number and tool platform details.

TABLE II. PROPERTIES OF THE CARVING TOOLS

Name	License	Version	Platform
MP4 -karver	Proprietary	v.1.1	Windows
PhotoRec	Open source	v 7.1	Multi
blade forensic	Proprietary	v1.11	Windows

#### C. Implementation Overview

The main aim of this research is to develop MP4-Karver tool for the carving of MP4 videos, repairing corrupted videos and extracting frames from videos. C-Sharp (C#) programming language is used for developing the proposed MP4-Karver tool. The hardware setup for this tool was Intel® Core i7 CPU and 4GB of physical memory with windows 7.

Figure 1 presents the steps of proposed MP4-Karver tool for recovering and repairing videos. Disk images are computer files that comprise two types, the contents and structure of whole data in the storage devices. Reading MP4 video in the disk image files, by using streaming files of MP4 video and this contains hexadecimal value. Hexadecimal format pattern is used by searching offsets to find the location of all the headers of the video file signature.

For the recovery MP4 videos, the signature of MP4 video header is first to be identified. The file header is about 4bytes size shown Figure 2. In this figure, the offset of 4 bytes (hex: 0x 66, 74, 79, 70) ftyp header signature of the MP4 video can be seen.

After verifying the header files of MP4 video using that pattern signature, it is possible to extract MP4 video data. Then the videos are recovered and the corrupted videos are passed to the second part of the method to be repaired. The repairing corrupted MP4 video files can be solved by taking advantage the structure of the video.

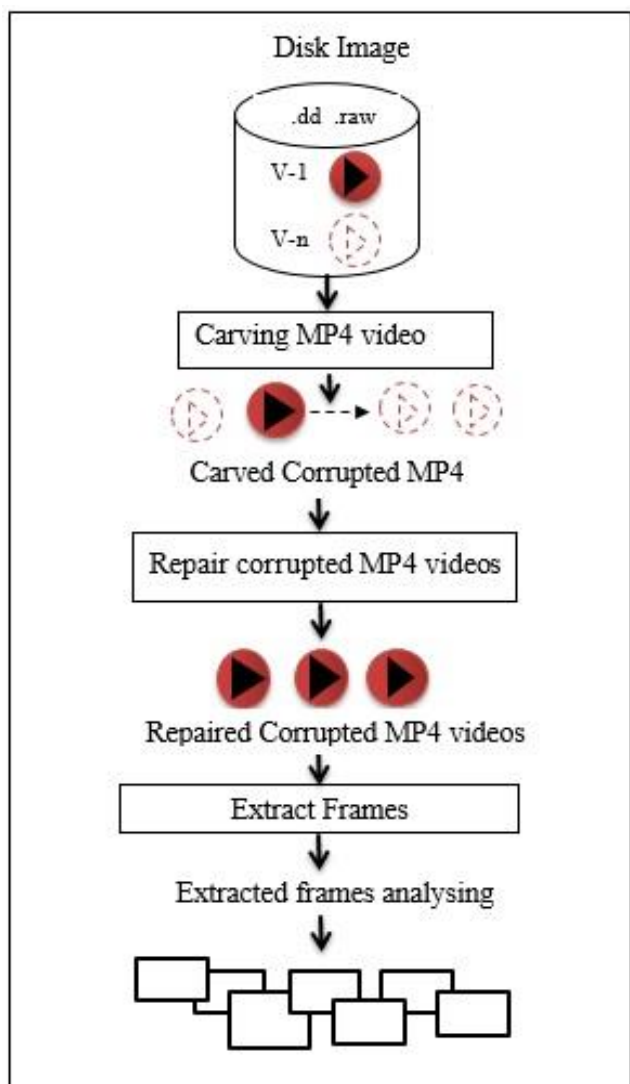


Fig. 1. Methodology of Proposed MP4-Karver Tool

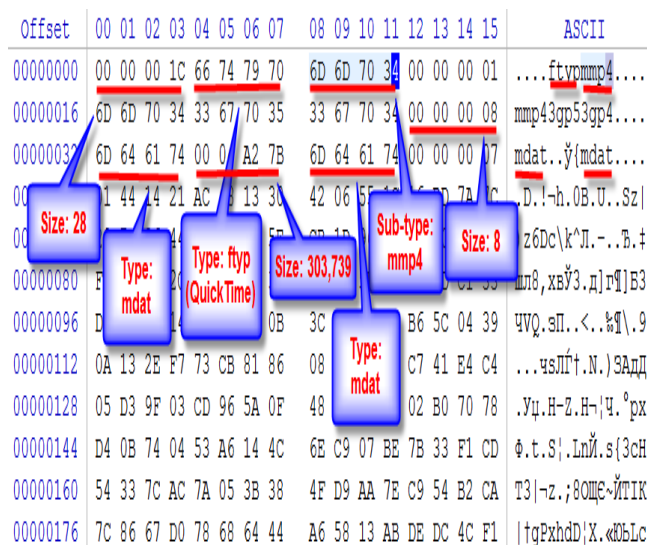


Fig. 2. MP4 Video Signature

Figure 3 shows a proper MP4 -format file structure. It begins with file type box (ftyp) and Movie box (moov), followed by a Mdat (Media Data Box). Normally, this 'mdat' would contain raw video data (H.264 NAL units). Also, this figure shows the corrupt video file which looks a like the normal structure however, it contains the 'ftyp' and 'mdat' atoms only. Therefore, the file can be repaired simply by skipping over the initial 'ftyp' and 'moov' atoms, and the header for the 'mdat' atom, and copying the remaining data into a new file. Now, the videos are fully recovered and repaired.

MP4 video after recovery and repaired videos then are extracted the video frames for automatically screening. A video file is actually a combination of several still images which are called frames. Extracting frame starts with reading number frames of video contains which are still images and can be captured. Then, captured frames are converted into image formats jpg.

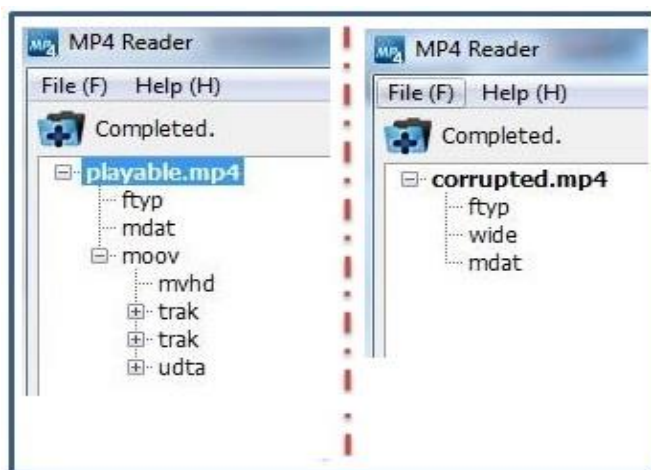


Fig. 3. Comparing Normal and Corrupted MP4 Structure

#### IV. EXPERIMENT AND RESULTS

To evaluate the performance of the proposed MP4-Karver tool for video carving and repairing corrupted videos, we use three data sets, i.e. 1) DFRWS 2) BCDS and 3) Level. These data sets contain corrupted and partial videos, which can be seen in Table 1. The experimental results of the total video carved by the proposed tool and the existing tools are given in Table 3.

TABLE III. TOTAL MP4 VIDEO CARVED

DATA SETS	Total Videos	File Carving Tools	Total Carved
DFWRS2007.img	8	MP4-karver	8
		Photorec	8
		Blade forensic	1
bcds.raw	3	MP4-karver	3
		Photorec	2
		Blade forensic	2
L3_Video.dd	1	MP4-karver	1
		Photorec	1
		Blade forensic	1

We tested the tools and the results can be seen in Table 3. It can be seen that for the BFRWS data set, 100% videos are

recovered by MP4-Karver and Photorec tools. The Blade Forensic tool recovers only one video. For BCDS data set in Table 3, it shows that, total of three videos are recovered by MP4-Karver tool. The Blade Forensic and Photorec tools recover only two videos. However, Photorec recovered and Blade Forensic tools, two videos recovered are playable. Whereas Level data set, in Table 3 one video is recovered by the MP4-Karver tool. The Blade Forensic and Photorec tools also one video is recovered.

After the recovery of the videos, they may be playable or corrupted. According to table 4, In DFWRS dataset the recovered videos after repairing by MP4-Karver are all playable while Photorec recovered only one playable. On the other side Blade Forensic recovered only one video and that is playable. Once the videos are recovered, they could be playable or corrupted. The recovered videos by MP4-Karver is playable while Photorec and Blade Forensic recovered video is not playable. The rest videos of video recovered are partial videos mean not complete and playable video. BCDS data set, three videos are recovered through using MP4-Karver tool and 2 video is playable and one is partial is not playable. In the Level data set, one video is recovered by Photorec and Blade Forensic tools and that video is partial not playable.

TABLE IV. REPAIRED AND PLAYABLE

DATA SETS	File Carving Tools	Playable-Complete	Not playable	Partial Video
DFWRS2007.img	MP4-karver	8	0	0
	Photorec	1	7	0
	Blade forensic	1	7	0
bcds.raw	MP4-karver	2	0	1
	Photorec	2	0	0
	Blade forensic	2	0	0
L3_Video.dd	MP4-karver	1	0	0
	Photorec	0	0	1
	Blade forensic	0	0	1

The percentage improvement of the MP4-Karver over existing tool is given in Table 5. The percentage improvement is calculated as follows:

$$P_{imMK}(E) = \frac{VCE - VCM}{|Videos\ in\ a\ Data\ Set|} * 100 \quad (1)$$

Where  $P_{imMK}$  is the percentage improvement of MP4-Karver over existing (E) tool. VCE is the video carved by existing tools and VCM is the video carved by MP4-Karver tool. The obtained value from VCE-VCM is divided by the total number of videos in a data set. To get the percentage 100 is multiplied. In next step the average of percentage improvement of MP4-Karver over existing tools is calculated by using the equation:

$$AP_{imMK}(E) = \frac{\sum_{i=0}^{|Data\ sets|} P_{imMK}(E)_i}{|Data\ Sets|} \quad (2)$$

The Equation 2 sums up all the values of Equation 1 obtained for all the data sets and divided it by the total number of the data sets.

TABLE V. AVERAGE IMPROVEMENT OF MP4-KARVER TOOL OVER EXISTING TOOLS

DATA SETS	File Carving Tools	Percentage Improvement
DFWRS2007.img	Photorec	88%
	Blade forensic	88%
bcds.raw	Photorec	0%
	Blade forensic	0%
L3_Video.dd	Photorec	100%
	Blade forensic	100%
Average	-	63%

Subsequently, the calculation of average improvement of MP4-Karver tool as compared to the existing tools. The Table 3 & 4 results obtained from it also used to view and presented a graphical way. Figure 4 shows the total of carved videos from data sets and by comparing the file carving tools. Also, Figure 5 demonstrates the result of repairing corrupted videos from a data set by using carving tools.

Next part highlights the results of the last step of the MP4-Karver video method of extraction frames. In Table 6 and Figure 6, the number of extracting frames from video recovered is presented, MP4-Karver tool is extracted 3522 frames from BFRWS data set where from BCDS data set extracted 19314 frames and the Level data set, and 262 frames are extracted. The Photorec and Blade Forensic tools they do not have the feature to extract frames from videos.

TABLE VI. RESULT OF VIDEO FRAMES EXTRACTED

DATA SETS	Total Videos	File Carving Tools	Total Extracted Frames
DFWRS2007.img	8	MP4-Karver	3,532
		photorec	0
		Blade forensic	0
bcds.raw	3	MP4-Karver	19,313
		photorec	0
		Blade forensic	0
L3_Video.dd	1	MP4-Karver	262
		photorec	0
		Blade forensic	0

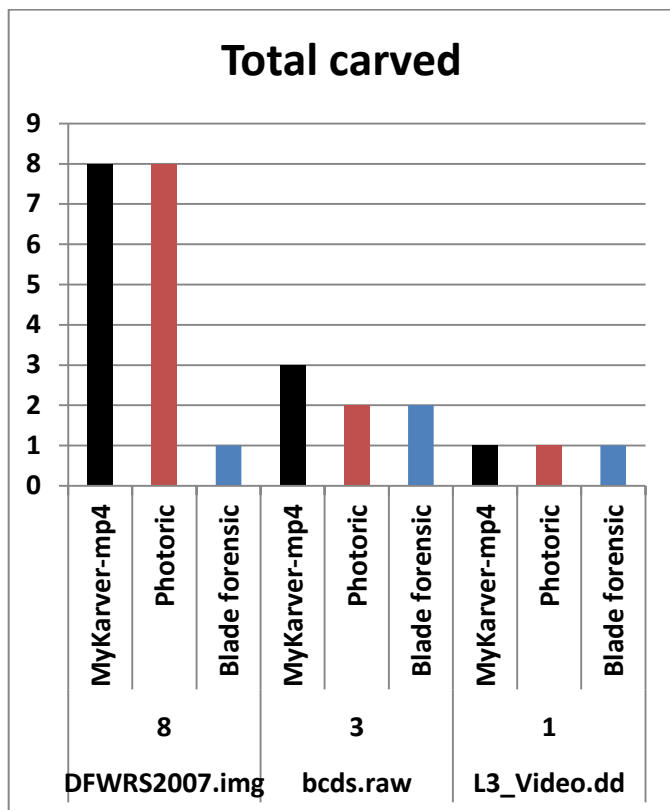


Fig. 4. Total video Carved

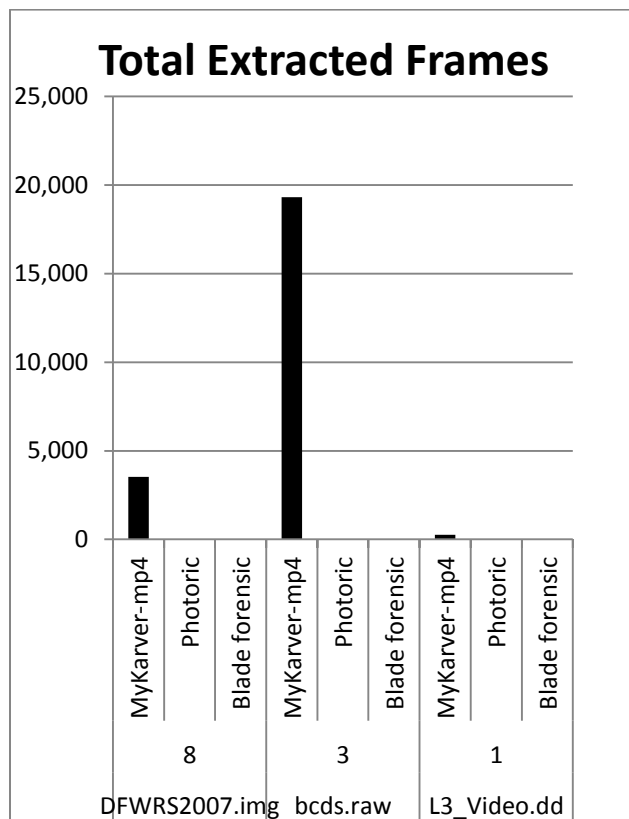


Fig. 6. Graphical representation of MP4 video frames

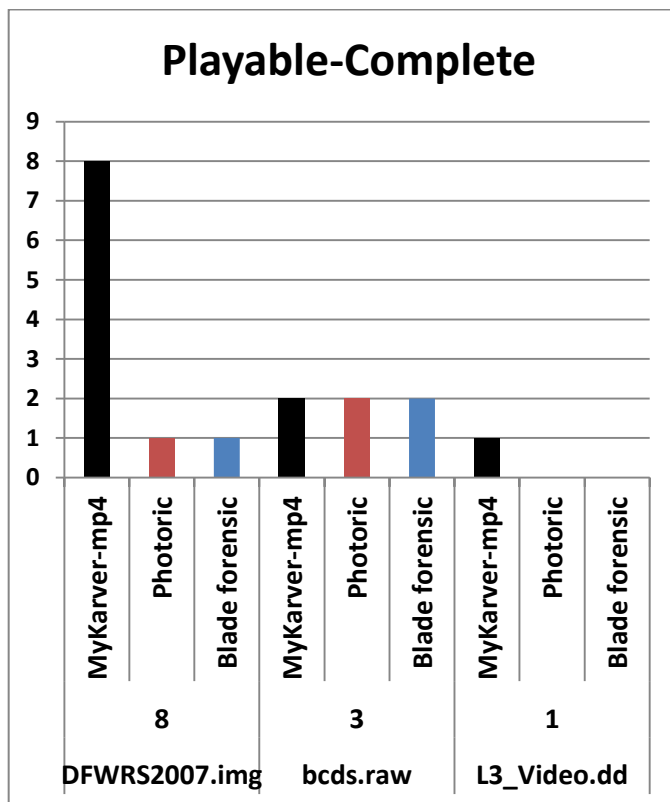


Fig. 5. Total video repaired

## V. CONCLUSION

This paper proposes a new MP4-Karver tool for video carving and repairing corrupted videos. The proposed MP4-Karver tool guarantees the efficient video restoration and repairing if it's corrupted or damaged and making complete playable videos and then frames can be extracted from videos. Moreover, MP4-Karver tool has the ability of restoring videos, if these are corrupted and overwritten or non-overwritten video files because many existing file-based techniques such as, Scalpel's method, Garfienkel, Bit fragment gap carving, smart carving and Frame based recovery cannot restore overwritten video files.

Most of the previous methods have been used to recover video files by searching for Metadata of the files, but the proposed tool recovers video file by taking advantage of the file structure and repairing corrupted video files without being affected by the corrupted parts. It is evaluated that the video restoration and repairing rate of MP4-Karver tool is higher than existing tools, i.e. Photorec and Blade forensic.

This research recommends doing further research on this tool and will add automatic video analysis for illegal activities detection such as child sexual graphic and also add enhancement, clarification and enlargement of the video in order to increase the filtering of evidence correctly to solve a criminal case.

## ACKNOWLEDGMENT

This research is supported under the Graduate Research Grants (GIPS) vote no. U307, Unisersiti Tun Onn Malaysia.



REFERENCES

- [1] Deshmukh, A., & Desai, M. "A Survey on Corrupted Video Recovery Using CODEC Specifications", International Journal of Science and Research (IJSR), Volume 3 Issue 12, December 2014.
- [2] G. H. Na, K. S. Shim, K. W. Moon, S. G. Kong, E. S. Kim, and J. Lee, "Frame-based recovery of corrupted video files using video codec specifications," IEEE Trans. Image Process., vol. 23, no. 2, pp. 517–526, 2014.
- [3] S. L. Garfinkel, "Carving contiguous and fragmented files with fast object validation," Digit. Investig., vol. 4, no. SUPPL., pp. 2–12, 2007.
- [4] B. Yoo, J. Park, and S. Lim, "A study on multimedia file carving method," 2011.
- [5] S. Hand, Z. Lin, G. Gu, and B. Thuraisingham, "Bin-Carver: Automatic recovery of binary executable files," Digit. Investig., vol. 9, pp. S108–S117, 2012.
- [6] V. Shoba and K. Gunasekaran, "An Integrated Approach for Recovering Corrupted Video Files using MPEG-4 Visual Specifications," vol. 3, no. 1, pp. 249–254, 2015.
- [7] Yannikos and Ashraf, "Automating Video File Carving and Content Identification," Adv. Digit. Forensics IX, pp. 195–212, 2013.
- [8] K. M. Mohamad, T. Herawan, and M. M. Deris, "Detecting JFIF Header using FORHEADER 1," vol. 5, no. 4, pp. 17–26, 2011.
- [9] S. Garfinkel, A. Nelson, D. White, and V. Roussev, "Using purpose-built functions and block hashes to enable small block and sub-file forensics," Digit. Investig., vol. 7, pp. S13–S23, 2010.
- [10] W. Qiu, R. Zhu, J. Guo, X. Tang, B. Liu, and Z. Huang, "A New Approach to Multimedia Files Carving," 2014 IEEE Int. Conf. Bioinforma. Bioeng., pp. 105–110, 2014.
- [11] Na, G. H., Shim, K. S., Moon, K. W., Kong, S. G., Kim, E. S., & Lee, J. (2014). "Frame-based recovery of corrupted video files using video codec specifications". Image Processing, IEEE Transactions on, 23(2), 517-526.
- [12] Metz, J., & Mora, R. (2007). Analysis of 2006 DFRWS forensic carving challenge. 2007-8-28). [http://sandbox.dfrws.org/2007/metz/dfrws2007\\_carving\\_challenge.pdf](http://sandbox.dfrws.org/2007/metz/dfrws2007_carving_challenge.pdf).
- [13] Thakre, C., Thakare, M., & Dhande, S. A. "Efficient Design of Frame Based Recovery Technique for Corrupted Video Files". INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING Vol. 3, Issue 6, June 2015
- [14] Avinash Deshmukh, Manisha Desai. "A Survey on Corrupted Video Recovery Using CODEC Specifications". Volume 3 Issue 12, December 2014.

# Competitive Representation Based Classification Using Facial Noise Detection

Tao Liu

Chongqing Key Laboratory of Computational Intelligence  
College of Computer Science and Technology, Chongqing  
University of Posts and Telecommunications  
Chongqing, China

Ying Liu

Chongqing Key Laboratory of Computational Intelligence  
College of Computer Science and Technology, Chongqing  
University of Posts and Telecommunications  
Chongqing, China

Cong Li

Chongqing Key Laboratory of Computational Intelligence  
College of Computer Science and Technology, Chongqing  
University of Posts and Telecommunications  
Chongqing, China

Chao Li

Chongqing Key Laboratory of Computational Intelligence  
College of Computer Science and Technology, Chongqing  
University of Posts and Telecommunications  
Chongqing, China

**Abstract**—Linear representation based face recognition is hotly studied in recent years. Competitive representation classification is a linear representation based method which uses the most competitive training samples to sparsely represent a probe. However, possible noises on a test face image can bias the representation results. In this paper we propose a facial noise detection method to remove noises in the test image during the competitive representation. We compare the proposed method with others on AR, Extended Yale B, ORL, FERET, and LFW databases and the experimental results show the good performance of our method.

**Keywords**—face recognition; sparse representation; biometrics; noise detection

## I. INTRODUCTION

Face recognition (FR) technology has been attracting many researchers to study so far [1,2]. In recent years, a lot of FR methods using linear representation are proposed [3,4]. There are at least two obvious merits of linear representation based methods: First, a test sample synchronously matches with all training samples from gallery set, which guarantees high recognition accuracy. Second, it is verified that the recognition results are not sensitive to the type of feature, which means very simple feature such as down-sampled images can serve well. In a linear representation based FR method, an assumption is made that samples from the same class are distributed in a same subspace, which means a given test image  $\mathbf{y}$  can be linearly represented by a training set  $\mathbf{A}$  as follows:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \quad (1)$$

where  $\mathbf{x}$  is a coding vector and in this model images are transformed from matrix into a vector. In classical biometric methods a test sample needs to match each training sample one at a time, but we can see from (1) that the correlation between the test sample and all training samples is built in one linear representation model. It is quite common that the training set matrix  $\mathbf{A}$  is over-complete, so that a sparse representation scheme is proposed to enforce only a few training samples to

respond in the representation model. One way to implement the sparse representation scheme is to introduce a  $l_0$ -norm constraint on a coding vector, which is formalized as:

$$\min \|\mathbf{x}\|_0 \quad s.t. \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \varepsilon \quad (2)$$

where  $\|\cdot\|_0$  denotes the  $l_0$ -norm, which counts the number of nonzero entries of the coding vector and  $\varepsilon$  is a small error tolerance. Since to solve the model (2) is a NP-hard problem, in [5] a new model was built as

$$\min \|\mathbf{x}\|_1 \quad s.t. \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \varepsilon \quad (3)$$

where  $\|\cdot\|_1$  is the  $l_1$ -norm of a vector, which sums up all absolute values of entries in the vector. It is proved that under some conditions the solution of (3) is a good approximation of (2). To optimize (3), a further equal transformation is made to link the sparse representation model to the Lasso regression, which is

$$\min \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 + \lambda \|\mathbf{x}\|_1 \quad (4)$$

where  $\lambda$  is used to balance the  $l_1$ -norm based regularization term and the  $l_2$ -norm based fidelity term.

Many follow-up studies improved the performance of the recognition using (3) in their sparse representation FR methods. In [6], robust sparse coding (RSC) methods are proposed which are robust to various kinds of noises. In RSC, the fidelity term is replaced with a sigmoid-like function, which makes the representation less sensitive to outlier pixels. In [7], a sparse representation model is proposed based on the maximum correntropy criterion, which is also a method using a more robust function to replace the  $l_2$ -norm based fidelity term in (4). But, furthermore, a new technique called half-quadratic framework is proposed in this work, which improves the sparse representation in terms of both error correction and error detection [8].

Some researchers argue that without imposing sparsity on the coding vector, one can still obtain good classification accuracy. In

[9], linear regression models are used to represent a probe one class at a time. And their later work uses Huber estimator to achieve more robust regression against different levels illumination changes. In [10], authors suggest that to use ridge regression instead of Lasso regression model is good for the case of Gaussian distributed noise and introduce the collaborative representation based classification (CRC) with the non-sparse  $l_2$ -norm to regularize the representation coefficients. However according to the investigation in [11] to restrict the number of training samples in the collaborative representation can increase the performance of CRC further. Inspired by this study, a two-phase collaborative representation method is proposed, in which the first phase is to choose a subset of training samples which is close to the test sample and the second phase is to conduct CRC on the chosen training samples. The two-phase collaborative representation can be considered as another way to realize sparse representation in the sense that only a few training samples are involved in final representation. This kind of sparse representation scheme is called supervised sparse representation because the sparsity comes from the supervision in representation itself not  $l_1$ -norm based regularization. In [12], experiments are conducted to show that to use several rounds of representation to subtly choose the training set is better than one-shoot training sample picking. It is not surprising that multiple rounds of training sample picking is quite time-consuming and however a fast calculation method, called competitive sparse representation classification (CSRC), is proposed in [13]. The method only deletes the lowest competitive samples in each round of representation and uses a fast algorithm to avoid repeatedly calculating the matrix inverse in each round.

However the supervised sparse representation has not considered the possible massive noises on a probe such as scarf occlusion so far. The noise could harm the procedure of picking the competitive training sample subset so that the chosen samples show the appearance of bias in favor of the occlusion. In this paper, we propose a new competitive sparse representation classification using facial noise detection (CSRC-FND). We do not only employ the multiple round of representation to select the most competitive subset of training samples, but also to delete the possible outlier pixels.

The paper is organized as follows: In section 2, we first introduce the competitive sparse representation classification method and then we propose our CSRC-FND method. In section 3, an analysis of the proposed method is given by illustrating the difference between CSRC and CSRC-FND. We conduct several experiments in section 4. And the conclusion is made in Section 5.

## II. THE PROPOSED METHOD

### A. Competitive representation classification

We first briefly introduce the competitive representation classification method.  $\mathbf{A} \in R^{m \times n}$  denotes a subset of training samples, where each column is an training sample with  $m$  pixels and  $n$  is the number of training samples in the subset. The procedure of competitive representation is multiple rounds of representation of a probe sample  $\mathbf{y} \in R^m$  and in each round

of representation the least competitive training samples are removed from the subset  $\mathbf{A}$ . For the first round,  $n$  equals the total number of training samples. To implement the competitive representation, the following ridge regression model is used:

$$\mathbf{x} = \arg \min_x \{ \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_2^2 \} \quad (5)$$

where  $\lambda$  balances the regularization term and the fidelity term. The coding vector  $\mathbf{x}$  has the close form solution:

$$\mathbf{x} = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{y} \quad (6)$$

where  $\mathbf{I}$  is an identity matrix. In each round, the coding vector  $\mathbf{x}$  is calculated according to the subset  $\mathbf{A}$ . Then we need to sort the absolute value of entries of  $\mathbf{x}$  and remove the samples associating with the smallest the absolute value (the number may be more than 1) from the subset  $\mathbf{A}$ . We repeatedly do the competitive representation until the subset  $\mathbf{A}$  is small enough.

In the final decision, (5) is used to find the coding vector again. Then the decision is made by

$$\begin{aligned} \text{ID}(i) &= \arg \min_i d_i(\mathbf{y}) = \arg \min_i \|\mathbf{y} - \hat{\mathbf{y}}_i\|_2 \\ \hat{\mathbf{y}}_i &= \sum_j a_i^j x_i^j \\ i &= 1, 2, \dots, c \end{aligned} \quad (7)$$

where  $c$  is the total number of classes,  $\hat{\mathbf{y}}_i$  is the prediction of the test sample by the  $i$ th class,  $a_i^j$  is a training sample of the  $i$ th class in the final subset and  $x_i^j$  is the corresponding coefficient in the coding vector.

### B. New method using facial noise detection

The competitive representation classification method performs very well in the cases where noises on test samples are not strong. However in some applications of FR faces can be contaminated by severe variations such as massive partial occlusion and pixel corruptions. These noises can make the procedure of competitive representation bias, which means the chosen training samples show preference for matching the noise so that the competitiveness of the genuine samples would be suppressed. Inspired by supervised sparse representation itself, we consider to combine the noise detection with the competitive representation procedure. Hence in each round of the competitive representation of a test sample we not only pick training samples with high competitiveness but also to identify some possible contaminated pixels. To facilitate the presentation of our method, a unified model is used to implement the idea given as:

$$\arg \min_x \|\text{Diag}(\mathbf{w})(\mathbf{y} - \mathbf{A}\mathbf{x})\|_2^2 + l \|\text{Diag}(\mathbf{d})\mathbf{x}\|_2^2 \quad (8)$$

where  $\text{Diag}(\mathbf{g})$  is to transform a vector to a diagonal matrix. Using this model, we use index vector to identify active pixels and active samples.  $\mathbf{w} \hat{\mathbf{I}} R^m$  is the index vector to specify which pixels are used in the model and  $\mathbf{d} \hat{\mathbf{I}} R^n$  specifies which training samples participate the representation. The entries of  $\mathbf{w}$  and  $\mathbf{d}$  are binary:

$$w_i \text{ and } d_j = \begin{cases} 1 & \text{active} \\ 0 & \text{inactive} \end{cases} \quad (9)$$

To set the value of  $\mathbf{w}$  and  $\mathbf{d}$ , it is easy to specify the active training samples and active pixels. The model (8) can be seen as a weighted representation model with binary weights and  $\mathbf{w}$  is the weight for pixels and  $\mathbf{d}$  is the weight for training samples. A good feature of model (8) is that it has close form solution as follows:

$$\mathbf{x} = (\mathbf{A}^T \text{Diag}(\mathbf{w})\mathbf{A} + I \text{Diag}(\mathbf{d}))^{-1} \mathbf{A} \text{Diag}(\mathbf{w})\mathbf{y} \quad (10)$$

At the beginning, we set the initial value of  $\mathbf{w}$  and  $\mathbf{d}$  to vectors with all entries equal to one. Then we start the competitive representation. In each round of representation we calculate the results by (10) and then we find the entries (associated with  $d_i = 1$ ) in  $\mathbf{x}$  which have the least absolute values and set corresponding values in  $\mathbf{d}$  to 0 (the changed values will not ever change back again). At the same time we calculate the representation residual by  $\mathbf{e} = \text{Diag}(\mathbf{w})(\mathbf{y} - \mathbf{A}\mathbf{x})$ , and find the entries in  $\mathbf{e}$  (associated with  $w_i = 1$ ) whose absolute values are greater than a preset small positive value  $x$  and set corresponding values in  $\mathbf{w}$  to 0 (the values will not ever change back again). By doing so, we try to identify the contaminated pixels on the test sample. The reason why we can find out corrupted pixels is that clear pixels can be matched by the linear model well while contaminated pixels deviate the real grey scale, which causes a big absolute residual. In addition we limit the lowest percentage of number of pixels that are involved in the representation. That is to say the ratio of zero entries to all number of entries in  $\mathbf{w}$  should be no more than  $r$  (which is preset).

The stop condition of competitive representation is that there are only the  $k$  most competitive training samples are active in the representation. That is to say, if there are only  $k$  entries in  $\mathbf{d}$  are equal ones, we stop the representation phase and do final representation. At last, the decision is made as in CRC method:

$$\mathbf{x}^* = \text{Diag}(\mathbf{d})\mathbf{x} \quad (11)$$

where  $\mathbf{x}^*$  is the final active coding vector.

$$\text{ID}(i) = \arg \min_i \|\mathbf{y} - \mathbf{A}_i \mathbf{x}_i^*\|_2 / (\|\mathbf{x}_i^*\|_2 + \tau) \quad i = 1, 2, \dots, c \quad (12)$$

where  $\tau$  is used to prevent denominator equal to zero. So that the proposed method can be summarized as the following algorithm:

**Algorithm:** Competitive sparse representation classification using facial noise detection (CSRC-FND)

**Input:** a probe image  $\mathbf{y} \in R^n$ , the initial dictionary  $\mathbf{A} \in R^{m \times n}$ .

Initial value:  $\mathbf{w} = [1, 1, \dots, 1] \hat{I} R^m$  and  $\mathbf{d} = [1, 1, \dots, 1] \hat{I} R^n$

**Repeat**

Compute the coding vector  $\mathbf{x}$  according to (10);

Set the entries of  $\mathbf{d}$  associated with the least absolute values of  $\mathbf{x}$  to 0;

$$\text{if } \frac{\text{sum}(\mathbf{w})}{m} < r$$

Set the entries of  $\mathbf{w}$  associated with the entries of  $\mathbf{e}$  that  $e_j > x$  to 0,

$$j = 1, 2, \dots, m;$$

end;

**Until**  $\text{sum}(\mathbf{d}) \leq k$

Compute the coding vector  $\mathbf{x}$  according to (10);

Compute the final active training samples according to (11)

;

$$\text{Identify : } \text{ID}(i) = \arg \min_i \|\mathbf{y} - \mathbf{A}_i \mathbf{x}_i^*\|_2 / (\|\mathbf{x}_i^*\|_2 + \tau) \\ i = 1, 2, \dots, c$$

**Output:**  $i$  (the identity of  $\mathbf{y}$ ).

### III. EXPERIMENTS

In this section, the performance of CSRC-FND is validated by extensive experiments. All the experiments are carried out using MATLAB on 2014a on a desktop with 3.30GHz CPU and 16G RAM. The normalized images are used in all experiments.

#### A. Parameter Setting

The proposed method includes three parameters, i.e. the representation steps  $N$  which represents the iterations of algorithm, the number  $M$  of removed feature in every step, and the number  $S$  of removed sample in every step. The three parameters codetermine which features and samples can be retained to linearly represent the testing image. Generally, the fewer the number of the retained noise and the number of the retained interferential samples are, the higher recognition rate of our method is. Through many experiments and analysis, the proposed method can obtain a good performance when  $N = 20$ ,  $M = 0.03 * m$ , and  $S = 0.04 * n$ .

#### B. Face Recognition without Occlusion

In this subsection, the performance of CSRC-FND is shown in FR without occlusion, such as posture variations and expression changes. Moreover, we compare the proposed method with LRC [9], CRC [10], SRC without expansion version [5], and CESR [7]. In addition, all experiments are performed on three public databases, namely, AR [14], ORL [15], FERET [16], and LFW [17].

##### 1) AR Database

In the experiment, 2600 images of 50 male and 50 female are chosen from the AR database. Each subject contains 26 images that are divided into two sessions, i.e. session 1 and session 2. We use 7 images of session 1 and session 2 for training and testing (i.e. Fig. 1), respectively. Moreover, the resolution of the images is  $50 \times 40$ . The recognition results of CSRC-FND and the compared methods are shown in TABLE I. CSRC-FND obtains the highest recognition rate among the

several methods, i.e. 96.71%. CESR uses a robust loss function to reduce the effect of noise in representation, but its recognition rate is lower 5.42% than our method. In addition, the recognition rate of CRC which uses  $l_2$ -norm to loss function is also lower than CSRC-FND. That is, it is very

effective that CSRC-FND removes some invalid pixels and samples for improving the recognition rate. Moreover, the recognition rate of CSRC-FND is 76.14%.



Fig. 1. The part of images of one subject in AR. Top: these facial images are the first seven images in session 1. Bottom: these images are the first seven images in session 2

TABLE I. THE RATES OF THE SEVERAL METHODS ON THE AR DATABASES

Methods	LRC	CRC	SRC	CESR	CSRC-FND
Recognition rates	76.14%	92.71%	75.71%	91.29%	96.71%

2) ORL Database:

ORL database contains 40 subjects and per subject consists 10 images with expression and posture changes, (i.e. Fig. 2). Each subject is separated into two parts: the first five images are used for training and the last five images are used for testing. All used images are cropped to  $50 \times 40$ . The in

TABLE II lists the recognition results of LRC, CRC, SRC, CESR and CSRC-FND. CESR and CSRC-FND obtain the highest recognition rate among all methods, i.e. 91.50% and 91.00%, respectively. In addition, the recognition rates of LRC (88%), CRC, and SRC are 3%, 5%, and 1.5% lower than our methods, respectively.



Fig. 2. The images of the first subject in ORL

TABLE II. THE RATES OF THE SEVERAL METHODS ON THE ORL DATABASES

Methods	LRC	CRC	SRC	CESR	CSRC-FND
Recognition rates	88.00%	86.00%	89.50%	91.50%	91.00%

3) FERET Database

FERET database consist of 1400 images of 200 subjects, i.e. each subject has 7 images, i.e. Fig. 3. These images mainly include posture variations. The first two images of per individual are using for training set, and the rest five images are used for testing. These images are resized to  $40 \times 40$  in this experiment. From in TABLE III, we can observe that the recognition rates of CESR and CSRC-FND are very close, the

margin is only 0.1%, i.e. CESR is 65.90% and CSRC-FND 65.80%. LRC also obtains very high recognition rate, i.e. 64.10%. The performance of CRC is poor, the margin between it and our method is 17.20%. Therefore, CSRC-FND largely improves the performance of recognizing face image with posture variations by removing invalid features and samples. In addition, SRC obtains a good result, i.e. 61.50%.



Fig. 3. The seven images of the first subject in FERET

TABLE III. THE RATES OF THE SEVERAL METHODS ON THE FERET DATABASES

Methods	LRC	CRC	SRC	CESR	CSRC-FND
Recognition rates	64.10%	48.60%	61.50%	65.90%	65.80%

4) LFW Database

This database is a large-scale database of face photographs designed for unconstrained FR, which includes of pose variations, illumination variations, expression variations, misalignment variations and occlusion variations, and so on. Some examples are presented in Fig. 4. We gray the images in the LFW database and resize into  $100 \times 80$ . We choose a subset which consist of 143 subjects with no less than 11 images each individual from LFW. Training set consists of the



Fig. 4. A part of images in a specific subject from LFW

TABLE IV. THE RATES OF THE SEVERAL METHODS ON THE LFW DATABASES

Methods	LRC	CRC	SRC	CESR	CSRC-FND
Recognition rates	40.09%	55.14%	24.31%	44.64%	62.17%

### C. Face Recognition with Occlusion

In this subsection, the robustness of methods is tested by dealing with the complex problems, such as corrupted facial, and occluded facial. And LRC, CRC, SRC, and CESR are compared with our method in these experiments. And the resolution of images in these experiments is  $50 \times 40$ .

#### 1) FR with Pixel corruption

In this experiment, 1984 images of 31 individuals are chosen from the Extended Yale B database [18,19] and each individual has 64 images which are divided into 5 subsets according to different light intensities.

As in Fig. 5 (a), subset 1 includes the low light intensity and this subset is used for training, and subset 3 with high light intensity shown in Fig. 5 (b) is used for testing. Moreover, a certain percentage pixels in the testing are randomly chosen and replaced with random value in the range between 0 and the maximum value of pixel in the testing (Fig.5. (c)).

We list recognition results of LRC, CRC, SRC, CESR and CSRC-FND under different percentage of corruption in TABLE V. We observe that performance of CSRC-FND is more stable than the other several methods under different corruption percentages. The recognition rate of our method occurs change until corruption percentage equals to 50%. However, the rates of the other methods are reduced when corruption percentage exceed 10%. It could illustrate that CSRC-FND can detect accurately the noise in testing facial image, and removing interfere samples is favor of face recognition.

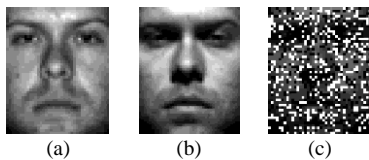


Fig. 5. Recognition under 40% random corruption. (a): An image in subset 1. (b): An image in subset 3. (c): A testing image with 40% random corruption

first ten images in each subject, and remaining images are used as testing. The recognition rates of the several methods are listed in TABLE IV. It is obvious that the margins between our method and the other several methods are very large the minima margin is 7.03%. Only is the recognition rate of CSRC-FND higher 60% among all methods. The recognition rates of LRC, CRC, SRC, and CESR are 40.09%, 55.14%, 24.31%, and 44.64%, respectively.

TABLE V. THE RATES (%) OF THE SEVERAL METHODS ON THE VERSUS DIFFERENT PERCENTAGE OF CORRPTUION

Corruption (%)	LRC	CRC	SRC	CESR	CSRC-FND
0	100.0	100.0	100.0	71.77	100.0
10	100.0	99.46	80.65	71.04	100.0
20	98.66	93.55	79.57	70.97	100.0
30	95.97	77.15	76.34	70.16	100.0
40	87.63	52.42	71.28	65.86	100.0
50	65.86	30.65	58.06	65.32	99.93
60	39.52	19.36	38.44	59.68	98.39
70	24.19	6.183	23.39	52.96	83.07

#### 2) FR with Block Occlusion

In this experiment, we validate the performance that CSRC-FND recognizes the images with block occlusion. And LRC, CRC, SRC, and CESR are tested as the comparisons. As in previous experiment, the subset 1 in Extended Yale B is used for training and subset 3 for testing. A middle block in the testing image is replaced by an unrelated image, i.e. Fig.6 (b). The results of LRC, CRC, SRC, CESR and CSRC-FND under different the occlusion percentage from 10% to 60% are listed in TABLE VI. From the table, we can observe obviously that 10% occlusion is more serious than 20%, 30% occlusion for recognition. This is due to the fact that the value of the noise with 10% occlusion has larger different with the value of the fidelity pixel than 20%, 30% occlusion. However, the recognition rate of our method still equals to 100%. In addition, the performance of CSRC-FND is very stable, the margin between 10% and 50% is only 3.76%. Even the occlusion percentage reaches at 60%, its recognition rate is equal to 73.66%. It is more robust than CESR with using robust loss function. The performance of the other methods is also nonideal, i.e. their performance is very sensitive to the occlusion percentage.

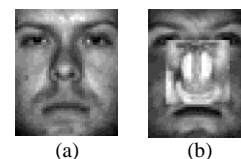


Fig. 6. Recognition under 30% block occlusion. (a) An image in subset 3. (b) A testing image with 30% block occlusion



TABLE VI. THE RATES (%) OF THE SEVERAL METHODS ON THE VERSUS DIFFERENT PERCENTAGE OF OCCLUSION

Occlusion (%)	LRC	CRC	SRC	CESR	CSRC-FND
10	68.28	37.10	54.83	68.55	100.0
20	75.27	57.80	59.95	63.98	100.0
30	77.69	47.31	63.44	64.25	100.0
40	70.67	43.01	60.21	55.91	99.46
50	68.55	43.01	54.84	52.15	96.24
60	47.85	30.11	40.05	37.90	73.66

#### IV. CONCLUSION

In this paper, a new competitive representation based FR method called CSRC-FND is proposed. In order to void the bias of choosing the training samples subset of high representation competitiveness, the noise detection method is employed to remove the pixels which are very likely to have been contaminated. To implement both the competitive representation and noise detection, a weighted regression model is presented which involves two binary weight vectors that one is used to identify the active samples and another is to specify the active pixels in repeated competitive representations. According to the experiments, the proposed method outperforms the competitive sparse representation classification and show promising performance. However, the noise on facial image is very various. It is hard for CSRC-FND to remove the noise on the image and inference samples in training set, accurately. Therefore, we will focus on this problem in the future.

#### REFERENCES

[1] L. Luo, J. Yang, J. Qian, and Y. Tai, "Nuclear-L1 norm joint regression for face reconstruction and recognition with mixed noise," *Pattern Recognition*, vol. 48, no. 12, pp. 3811-3824, 12, 2015.

[2] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *Acm Computing Surveys*, vol. 35, no. 4, pp. 399-459, Dec, 2003.

[3] Y. Xu, Q. Zhu, Z. Fan, D. Zhang, J. Mi, and Z. Lai, "Using the idea of the sparse representation to perform coarse-to-fine face recognition," *Information Sciences*, vol. 238, no. 0, pp. 138-148, 2013.

[4] J.-X. Mi, J.-X. Liu, and J. Wen, "New Robust Face Recognition Methods Based on Linear Regression," *Plos One*, vol. 7, no. 8, pp. e42461, 2012.

[5] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine*

*Intelligence*, *IEEE Transactions on*, vol. 31, no. 2, pp. 210-227, Feb, 2009.

[6] Y. Meng, Z. Lei, Y. Jian, and D. Zhang, "Regularized Robust Coding for Face Recognition," *Image Processing, IEEE Transactions on*, vol. 22, no. 5, pp. 1753-1766, 2013.

[7] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum Correntropy Criterion for Robust Face Recognition," *Ieee Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1561-1576, Aug, 2011.

[8] H. Ran, Z. Wei-Shi, T. Tieniu, and S. Zhenan, "Half-Quadratic-Based Iterative Minimization for Robust Sparse Representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 2, pp. 261-275, 2014.

[9] I. Naseem, R. Togneri, and M. Bennamoun, "Linear Regression for Face Recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 11, pp. 2106-2112, 2010.

[10] D. Zhang, Y. Meng, and F. Xiangchu, "Sparse representation or collaborative representation: Which helps face recognition?," pp. 471-478.

[11] Y. Xu, D. Zhang, J. Yang, and J. Y. Yang, "A Two-Phase Test Sample Sparse Representation Method for Use With Face Recognition," *Ieee Transactions on Circuits And Systems for Video Technology*, vol. 21, no. 9, pp. 1255-1262, Sep, 2011.

[12] J.-X. Mi, "Face image recognition via collaborative representation on selected training samples," *Optik-International Journal for Light and Electron Optics*, vol. 124, no. 18, pp. 3310-3313, 2013.

[13] Y. Liu, J.-X. Mi, C. Li, and C. Li, "Competitive Sparse Representation Classification for Face Recognition," *International Journal of Advanced Computer Science & Applications*, vol. 1, no. 6, pp. 1-7.

[14] A. M. Martinez, "The AR face database," *CVC Technical Report*, vol. 24, 1998.

[15] F. S. Samaria, and A. C. Harter, "Parameterisation of a stochastic model for human face identification." pp. 138-142.

[16] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295-306, 1998.

[17] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[18] A. S. Georghiadis, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 6, pp. 643-660, 2001.

[19] K. C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans Pattern Anal Mach Intell*, vol. 27, no. 5, pp. 684-98, May, 2005.

# Energy Efficient Routing Protocol for Maximizing the Lifetime in Wsns Using Ant Colony Algorithm and Artificial Immune System

Safaa Khudair Leabi  
Mechatronics Engineering Dept.  
Middle Technical University  
Baghdad, Iraq

Turki Younis Abdalla  
Computer Engineering Dept.  
University of Basrah  
Basrah, Iraq

**Abstract**—Energy limitations have become fundamental challenge for designing wireless sensor networks. Network lifetime represent the most important and interested metric. Several attempts have been made for efficient utilization of energy in routing techniques. This paper proposes an energy efficient routing technique for maximizing the networks lifetime called swarm intelligence routing. This is achieved by using ant colony algorithm (ACO) and artificial immune system (AIS). AIS is used for solving packet LOOP problem and to control route direction. While ACO algorithm is used for determining optimum route for sending data packets. The proposed routing technique seeks for determining the optimum route from nodes towards base station so that energy exhaustion is balanced and lifetime is maximized. Proposed routing technique is compared with Dijkstra routing method. Results show significant increase in network lifetime of about 1.2567.

**Keywords**—ant colony algorithm; artificial immune system; adaptive routing; network lifetime; wireless sensor networks

## I. INTRODUCTION

Recent years advances show serious progress in wireless networking. The progress and growth in wireless communication technology have made WSNs attractive for multiple application areas, such as medical and health, security surveillance, habitat monitoring, military reconnaissance, disaster management, industrial automation, etc. [1-4]. The development of small and ubiquitous WSNs computing devices is ultimately required. WSNs are comprised of considerable number of limited capabilities sensor nodes with one or more high capability base stations. Each sensor node is a small embedded system, low-power, low-cost, multi-functional [3] Each sensor node performs several functions: sensing, data processing, and communication. Sensor nodes perform wireless communications with each other in order for delivering gathered data to base station. The development of ubiquitous, inexpensive, small and low-power computing devices became available through miniaturization technologies [3]. Due to this, using multi-hop communication helps to reduce transmission distance as well as increasing network lifetime. Every node consists of four parts: a processor, sensor, transceiver, and battery. Nodes involve bounded power source with abilities of sensing, datum processing along with communication. The onboard sensors collect datum about the environment through event driven or continuous working

mode. The gathered datum may be temperature, pressure, acoustic, pictures, videos, etc. The gathered datum is then transferred across the network in order to form a global monitoring view for objects [5,6].

Since bounded energy source is involved, energy exhaustion is the most important metric for WSNs. In order for maximizing networks lifetime, energy exhaustion must be well managed [7,8]. Balancing energy exhaustion refers to the major problem in characterizing WSNs. Network lifetime might reduce significantly if the energy exhaustion is not balanced, and may lead to network partition quickly. Several techniques can be used for maximizing network lifetime, in which one of the important technique is network layer routing. Generally, in network layer routing algorithm, choosing the best route between nodes and base station represent the main objective of routing algorithms. If same path be chosen for all new communication by taking the benefit of fast transmission at the expense of battery energy exhaustion, sensor nodes of this path would drain its energy quickly and may cause network partition.

In this paper, an improved routing technique is proposed. The main goal of the suggested routing technique is balancing energy exhaustion along with maximizing network lifetime. The suggested routing technique is established by combining ant colony algorithm and artificial immune system so as for finding optimum path from sensor nodes towards sink.

The rest of the paper is organized as follows: related work is stated in section II. The suggested routing method is presented in section III. Simulation settings and configuration is presented in section IV. Section V describe simulation results and discussion. Conclusion is depicted in section VI.

## II. RELATED WORK

The primary consideration upon WSNs refer to the design of energy efficient system. WSNs routing protocols have been widely used for minimizing energy exhaustion along with maximizing overall network lifetime. Routing techniques is responsible for determining the optimum route from nodes towards base station. The good routing protocol involves determining the optimum path with the investigating of minimum energy exhaustion. Hence, network lifetime is maximized. Maximizing the lifetime of the network has gained significant interest in recent years.

The work proposed by Aniket et al. [9] presents a bio inspired ACO optimization for optimum route identification. The proposed method is achieved by three steps. First, the base station made aware about network topology. Second, the base station find the optimal path using ant colony optimization. Last, the optimal path is used by sensor nodes to communicate with other neighbor nodes and the base station. Simulation results show an increase in throughput and an increase in network lifetime.

Fatemeh [10] proposed ACO routing technique with colored pheromones with clustering method in order to satisfy QoS metrics. Four types of classes have considered in the proposed method. Each class identified by color pheromone. These types provide some routing metrics. Node energy, bandwidth, energy exhaustion, and end-to-end delay are used as routing metrics for the proposed method. Simulation is carried out in GloMoSim simulator. The proposed method is compared with EAQR and AntSensNet methods. Simulation demonstrates the performance of suggested method is better than other two methods with respect to lifetime, throughput, packet delivery ration, and delay.

The work proposed by Ahmed et al. [11] presented a self-optimization method for optimize sensor node energy to investigate energy consumption balance by using ant colony optimization to enhance the route with the best path. Several routing metric are used in this method. These metrics are residual energy, hop count, and average energy of route and network. Simulation is carried out in QualNet simulator. The proposed method is compared with EEABR method. Simulation results demonstrate an increase in the lifetime and reduction in energy consumption.

Bharathi et al. [12] proposed a reversed game theory for aggregation nodes selection with ACO routing algorithm. The suggested algorithm first selects a node for aggregation, which is a cluster head, then ant colony algorithm finds a route toward base station through cluster heads. Energy consumption is used as routing metric. Simulation is carried out in MATLAB. The proposed method is compared with LEACH algorithm. Simulation demonstrates an increase in lifetime, low energy consumption, high throughput, low delay, and high packet delivery ratio.

The work proposed by Lianyu et al. [13] presented a routing technique depending on artificial ant colony system. The suggested routing technique is collaborated with the direct diffusion routing. The proposed method consider the network as two categories, main nodes and other nodes. Main nodes are nodes that are located nearest to base station, which exhausts their energy faster. The proposed method try to maintain energy consumption of master nodes. Simulation results demonstrate that the presented method has average energy consumption of master nodes and maximize network lifetime.

Wenjing et al. [14] presented a comprehensive routing method by using ant colony algorithm. The proposed routing method uses four routing metric, remaining energy, square distance, path pheromone strength with nodes reputation. Simulation is carried out in NS2. The proposed method is compared with energy and activity aware (EAR) routing

protocol. Simulation results show an increase in lifetime and decrease in packet loss rate.

Wang et al. [15] presented LEACH clustering routing algorithm by using ant colony algorithm for WSNs. The proposed algorithm seeks for finding optimum path between cluster heads and the sink. Remaining energy and cluster head distance to sink are used as routing metrics. The proposed method is compared with LEACH routing protocol. Simulations demonstrate that suggested method has decreased energy exhaustion along with increasing network lifetime for about 30%.

The work proposed by Zhen et al. [16] presented ant colony system routing algorithm by using energy prediction. In the proposed routing protocol, remaining energy is used as the routing metric. When a node wants send data to base station, ant colony systems determine the route with optimum energy consumption, and learning mechanism involve predicting power exhaustion for neighboring nodes after node selects optimum one for route. Comparison with minimum spanning tree algorithm, least energy tree algorithm, and Dijkstra algorithm is presented. Numerical experiments show that the proposed routing algorithm gained lifetime best than the other three routing algorithms along with keeping power consumption in low level.

The work proposed by Yaofeng et al. [17] presented an energy and delay model algorithm depending on ACO algorithm (E&D Ants). The proposed technique try to minimize the delay for transferring a fixed number of data using an energy-constrained manner. Node energy and delay are used as routing metrics for the proposed algorithm. Simulation is carried out in OPNET. The proposed method is compared with AntNET algorithm. Simulation results show increase in lifetime, decrease in energy consumption, high throughput, and low packet loss.

Selcuk et al. [18] proposed an adaptive ACO routing algorithm. They have used the ant colony optimization to get a dynamic and reliable routing protocol. Node energy and number of neighbors nodes are the two metrics that is used in this routing protocol. The proposed routing protocol is implemented on a small sized router chip. The implementation is examined and efficiency demonstrated by simulation results, which are obtained by Proteus program. Results show an increase in network lifetime.

### III. THE PROPOSED ROUTING METHOD

The proposed routing technique involve ant colony algorithm and artificial immune system. The objective due to proposed routing protocol is balancing energy exhaustion so as for maximizing network lifetime. Ant colony algorithm is used for finding the optimum path from sensor nodes to sink. Artificial immune system is used to solve packet LOOP problem and to control route direction.

Sensor nodes are responsible for collecting data from its neighbors nodes. In this paper, time driven routing schedule is supposed. So that each node find the optimal rout for sending data packets to the sink in every time cycle. Using this scheduling for routing, the procedure of determining optimum route while sending data packets to the sink for all nodes will

repeated for each round. For the proposed routing protocols: (1) for a specified field random deployment is involved for whole nodes along with knowledge about their positions and their neighbors positions within its range and the position of the sink; (2) initial energy and maximum transmission range are identical for all sensor nodes.

Energy management efficiency is the most important WSNs design challenges; it gives a measure about WSN lifetime which perhaps the most serious metric owing to evaluating WSNs. The definition of net lifetime give a meaning for time from net turning on till first node exhausts its energy. The lifetime is the most challenging problem in WSN. Several techniques might be used to maximize network lifetime in which one of them is improving the routing protocols.

This paper suggests an improved routing technique for maximizing overall network lifetime. The suggested routing technique is achieved by combining ant colony algorithm and artificial immune system. Two metrics have been used, which are the remaining energy and shortest hop, for selecting optimum next hop node. The proposed routing protocol is responsible for selecting next hop of highest remaining energy and shortest hop to sink. Hence, energy exhaustion can be balanced along with maximized network lifetime.

The general structure for suggested routing technique is illustrated in figure 1. Suggested algorithm is working according to the following detail. When any sensor node like to send packet to the sink, the algorithm, firstly, find all its neighbors, and then by applying artificial immune negative selection algorithm, the algorithm classify good neighbor nodes (GNBR) and bad neighbor nodes. The good neighbor nodes will contribute in the final operation of finding the optimal path for sending or forwarding the packet. If there is no at least one good neighbor node, the algorithm response is failure because there is no path to send the packet. The operation of finding the optimal path involves ant colony algorithm with two metrics, which are the node remaining energy and shortest hop to the sink. The ant colony algorithm make a balance between these two metrics so as to find the optimal next node. This operation involve finding the probabilities of good neighbor nodes (GNBR) to be the next node depending on the probability function value of ACO algorithm. Probability function with regard to ACO algorithm has two parameters, which are  $\tau$  (pheromone concentration) and  $\eta$  (visibility function that is heuristic function). These two parameters are connected to our routing problem in which the pheromone concentration is considered as the node remaining energy (RE) and the visibility function is considered as the shortest hop (SH). The max-min ant system is considered which involves the pheromone is start with max and end with min for each city visit. The proposed algorithm calculates the probability values for all good neighbor nodes (GNBR). Then it, from the good neighbor nodes (GNBR), select the GNBR node of highest probability value to be the next optimal node and add it to OPEN list and flag it as the current node. The proposed algorithm then check this node if it is inside sink range. If so the algorithm ends and the optimal path is OPEN list. If no, the current node (the latest node added to OPEN list) is then uses the proposed algorithm to

find its optimum node to send packet (classifying GNBR nodes, finding their probability values, and selecting node of highest probability value). In other words, the algorithm is repeated for the new node that like forward the packets. The OPEN list is used in the algorithm to temporality store the nodes for the operation of finding the optimal path. OPEN list is used to store the elected nodes for finding the optimal path. The OPEN list is then stores the optimal path when the algorithm ends and the optimal path is found. Hence, the operation of the proposed algorithm is repeated for every node like send or forward data packets.

#### A. Artificial Immune System

The AIS has deduced from biological immune system of humans that defend the body against the threats. Naturally, the immune system is so complicated system and involve several functions. The master function is to defend against attacker cells. This is achieved by using two techniques, which are innate and acquired techniques. The master function takes the responsibility of classifying human cells into two categories which are self and non-self cells [19]. By using special type of defense, the immune system enforce non-self cells for some treatments that lead them to disintegration. The immune system has the ability of learning via mutation in order for distinguishing among external antigens, such as bacteria or viruses, and self cells of body.

Immune system framework involves negative selection process. The purpose of this algorithm is providing existence probability of self cells [20]. Processing of this algorithm involves some activities. First is ability of detecting strange antigens along with discarding reaction to self cells. By using random processing genetical arrangement, receptors be made during T-cells generation. Second, under sensory processing in thymus, receptors' T-cells that affect self-proteins subject for destroying along with allowing ones that not affect self-proteins for leaving thymus. The forward T-cells spread over the body for subjecting to immunity reactions. This processing can protect bodies from strange antigens. The algorithm artificially has been used in this thesis and the negative selection has been applied.

Forrest [21], in 1994, were firstly introduced the algorithm for detecting datum in computer systems when they handling viruses. This processing is carried out by generating two sets. The first set is self categories that give an indication about normal situation of that system. The second set is detectors that discover S-complement. Hence, datum have to be subjected to detectors set so as for recognizing them as self and non-self.

This section investigates the application of artificial immune in WSNs to solve packet LOOP problem and to control the direction of the route. Artificial immune system provides the ability for detecting danger neighbor nodes. This is done by classifying as good neighbor nodes (GNBR) and bad neighbor nodes (danger nodes). Artificial immune negative selection algorithm is used in the proposed routing method. Route direction control is crucial feature in the design of any routing algorithm. If this algorithm is not used, the packet may sent faraway from sink and encounter packet LOOP. Implementing route direction control will improve the

routing algorithm, ensure that the packet will not get a path faraway from sink, minimize energy exhaustion, and maximize network lifetime.

Good neighbor nodes (GNBR) will contribute to the optimal path finding, using the ant colony algorithm, while the bad neighbor (danger neighbors) will discarded from contributing to find the optimal path. The criteria used for classifying the good and bad neighbors nodes depends on the distance to sink metric. The artificial immune negative selection algorithm that used in the proposed routing algorithm is as listed below.

---

**Algorithm 1: Negative Selection Algorithm**

---

1. Inputs : S has the criteria to classify good neighbor nodes;  $S = \{S: s \in P | \text{distance to sink } D(s) < \text{distance to sink } D(n)\}$ ; where, n is the node want to send packets, s is a neighbor node, P is the set of all neighbors nodes for node n
  2. Output : D is the set of good neighbors nodes (GNBR)
  3. Repeat for all neighbors nodes in the set P for node n
    - a. Determine the affinity of each member of P with good neighbors criteria S
    - b. If neighbor node satisfies the criteria S, add it to set of good neighbors D
    - c. Else discard the neighbor node and define it as danger node for node n.
  4. Stop when all neighbor nodes for node n has been classified
- 

**B. Ant Colony Algorithm**

ACO algorithm is presented as swarm intelligence procedure along with involving some metaheuristic optimizations. ACO algorithm was first presented in 1992 by Dorigo in his PhD thesis [22]. Ant colony algorithms studies is derived from the observation of real ants' behavior. Its first involvement was for seeking for an optimum trajectory inside a diagram depending on ants' behavior for searching paths among their colony and food sources. Naturally, ants wandering at random, when discovering eating go back to their colony along with putting down pheromone lines. Then ants follow these lines instead random wandering. Pheromone is subjected to an evaporation. Subsequently, ants go after good path found among colony and eating.

ACO algorithm conception involves representing the problem being solved with demonstrated ants travelling around graph that mimic the behavior of natural ants. Ant colony algorithms are used as an improved design for novel algorithms for solving optimization and distributed control problems. The ant colony algorithm is outlined below [22].

---

**Algorithm 2: Ant Colony Optimization Algorithm**

---

Procedure ACO MetaHeuristic

1. While (cover all GNBR nodes)
  - a) GenerateSolutions();

$$Prob(k) = \frac{\tau^{\alpha}(k) * \eta^{\beta}(k)}{\sum \tau^{\alpha}(k) * \eta^{\beta}(k)} \quad (1)$$

- b) DaemonActions(); Select the node of highest probability value from set GNBR nodes and add it to OPEN list.
- c) PheromoneUpdate(); A Max-Min Ant System is used.

$$\tau_{new} = (1 - \rho)\tau_{current} + \Delta\tau \quad (2)$$

2. End while

End procedure

---

where, Prob(k) is the probability function,  $\tau(k)$  is the ants pheromone concentration,  $\eta(k)$  is the visibility function (heuristic function), k is GNBR member node,  $\alpha$  and  $\beta$  refer to ACO parameters that utilized for adjusting pheromone strength with heuristic magnitude for specified significance. The values of  $\alpha$  and  $\beta$  are restricted,  $\alpha \geq 0$ , and  $\beta \geq 1$ .

$\rho$  is the pheromone evaporation ratio and is expressed by the following equations.

$$\text{For Transmitting Packets, } \rho_{Tx} = E_{Tx}/E_{res} \quad (3)$$

$$\text{For Receiving Packets, } \rho_{Rx} = E_{Rx}/E_{res} \quad (4)$$

$E_{res}$  is the residual energy.

$$\Delta\tau = \begin{cases} \frac{Q}{L} & , \text{ if ant pass curve} \\ 0 & , \text{ otherwise} \end{cases} \quad (5)$$

Q is constant and L is the length of the path found by the ant.

This section investigates the application of ACO algorithm in WSNs for determining optimum route from nodes towards base station. ACO algorithm provide the ability for balancing energy exhaustion with the use of shortest distance to the sink. Balancing energy exhaustion leads for maximizing network lifetime. The suggested ACO algorithm not only balance energy exhaustion but also determines the optimal solution of node energy and node distance combination. This structure leads to the efficient optimum solution that ensure energy balance and maximum network lifetime.

For the proposed routing protocol,  $\tau$  is the ants pheromone concentration and is related to node energy,  $\eta$  is the visibility function and is related to the node shortest hop to sink SH(n). The visibility function is expressed in equation (6). Each GNBR node is considered as a city for the simulated ants. Ant visit all cities (GNBR nodes) at least one time. A Max-Min ACO arrangement has used for updating pheromone. Maximum and minimum ant system pheromone amounts suppose that all edges are initialized to  $\tau_{max}$  and reinitialized to  $\tau_{max}$  when nearing stagnation. For the proposed routing protocol, the nodes residual energy is considered as  $\tau_{max}$ .

$$\eta = \frac{1}{SH(n)} \quad (6)$$

where, n is a sensor node.

**IV. EXPERIMENTAL SETTINGS AND CONFIGURATION**

Simulation is carried out in MATLAB. Two topological areas are considered in this paper, which are A1 and A2. A 100 nodes are randomly scattered for every two topological areas. The topological areas A1 and A2 have the dimension of 100mx100m for area A1 and 200mx50m for area A2. One base station "Sink" has been used for each topological area.

The position of the sink is (90m,90m) for topological area A1, and is (180m,45m) for topological area A2. Every node operates with maximum transmission range equals to 30m. Every node has initial energy equals to 0.5J. A 200 bit packet length is used for simulation. The value of hop count limit (HCL) is equals to 10 and 15 for areas A1 and A2, respectively. Performance evaluation of suggested routing technique is tested for each of the two topological areas A1 and A2. The proposed routing technique utilized with first order radio model proposed by [23]. This model is shown in the following equations.

$$E_{TX}(pkt_{length}) = E_{elec} * pkt_{length} + E_{amp} * pkt_{length} * d^2 \quad (7)$$

$$E_{RX}(pkt_{length}) = E_{elec} * pkt_{length} \quad (8)$$

where,  $E_{TX}$  and  $E_{RX}$  are the energy exhaustion for transmitting and receiving respectively.  $pkt_{length}$  represents number of bits per packet.  $d$  represents distance between two communicating nodes.  $E_{elec}$  represents per bit energy exhaustion for broadcasting or receiving for electrical hardware.  $E_{amp}$  is the per bit per meter square energy exhaustion.  $E_{elec}$  and  $E_{amp}$  values that used for simulation are 50nJ/bit and 100pJ/bit/m<sup>2</sup>, respectively. For ant colony algorithm, magnitudes of ( $\alpha, \beta, \Delta\tau$ ) have chosen as (0.5, 2, zero), respectively.

## V. RESULTS AND DISCUSSION

Simulation is carried out for the two topological areas. Two routing protocols have been considered in the simulation, which are Dijkstra routing, and the proposed routing protocol. Number of alive nodes in each round has used to give the indication about the lifetime of the WSNs. A comparison has been made for the two routing techniques with reference to overall network lifetime beside other metrics. Network lifetime defined as period between network starting work till the first sensor node die or exhaust its energy.

Figure 2 depicts network lifetime for the two topological areas A1 and A2 in terms of number of still alive nodes in each round till network partition. From this figure, it can be seen that the proposed energy efficient swarm routing protocol is better than the Dijkstra routing. It shows an increase in network lifetime of about 1.2567 for area A1 and 1.9717 for area A2. Without using artificial immune system route direction control, the packets sent will encounter loops and may not successfully delivered to the sink causing the PDR to decrease to 0.877 for area A1 and 0.881 for area A2. Results show the improvement of the proposed energy efficient routing protocol in comparison with the Dijkstra routing. The proposed method shows the improvements in terms of balancing energy consumption, route direction control, solving packet LOOP, and maximizing overall network lifetime. Depending on the trace of experiments for searching optimal path, the proposed routing protocol changes the optimal path every round depending on the metrics used, remaining energy and shortest hop towards base station. This change in the path used to send data packets from any node to sink proves the balance in energy consumption and as a result maximizes network lifetime. A network partition feature has been activated for the simulation. Network partition is works out when any of the 100 deployed sensor nodes has not find a

neighbor nodes to send data packet. This is due to the dyed sensor nodes. Hence, simulation is stopped when network partition is occurred. Table 1 details the overall network lifetime, partition time, and PDR for the two methods and for the two topological areas A1 and A2.

Figure 3 illustrates the network average remaining energy in each round for the two topological areas A1 and A2 as a comparison between the proposed energy efficient routing protocol and Dijkstra routing. It is obvious from this figure, that the average remaining energy of the suggested energy efficient routing technique is less than Dijkstra routing, which gives the fact that Dijkstra routing has more nodes subjected to inefficient use by unbalanced energy exhaustion. Less remaining energy due proposed routing reflects the efficient use of nodes that provide energy consumption balance. This figure shows that Dijkstra routing method has some nodes exhaust its energy quickly due to the continuous usage of these nodes. This reflects the unbalanced energy consumption in the Dijkstra routing. This is due to the usage of the same path for sending data packets from sensor node to sink. According to the experiments trace for optimal path finding process, the proposed routing method changes the optimal path for each round depending on the remaining energy (RE) and shortest hop (SH) to sink metrics. So that the route used to send data packets from a sensor node to the sink, take different path in each round to deliver the data packets. In addition, tracing results show that without using the artificial immune route direction control, the packets sent will encounter loops and may not delivered to the sink. The proposed energy efficient routing technique results in energy exhaustion balance, unity PDR along with maximized overall lifetime network.

Figure 4 illustrates the average consumed energy in each round for the two topological areas A1 and A2. From this figure, consumed energy for proposed SI routing is higher than the Dijkstra routing. The key feature of the proposed SI routing protocol is investigating energy consumption balance and avoiding the continuous using of same nodes by using more hop nodes. This reflect the effectiveness of proposed swarm routing protocol for balancing energy exhaustion along with prolonging network lifetime and guarantying unity PDR.

Figure 5 depicts the Packet delivery ratio in each round for the two topological areas A1 and A2. From this figure the proposed SI routing shows unity PDR method along with higher network lifetime in comparison with Dijkstra routing which shows less network lifetime. This result give the power point for the proposed SI routing method which guarantying unity PDR along with higher network lifetime. Also this figure shows the importance of introducing the artificial immune system, in which without using artificial immune system for route direction control and solving packet LOOP problem, the PDR will decrease to 0.877 for area A1 and 0.881 for area A2.

Figure 6 depicts the maximum number of hops in each round for the two topological areas A1 and A2. This figure shows that proposed SI routing is better than Dijkstra routing, which change the path for sending packets continuously depending on the remaining energy and shortest hop metrics instead of using the same path as Dijkstra routing. Changing



transmission path leads to energy consumption balance by using more hops and avoid continuous using for some nodes. Using more hops helps to investigate energy consumption balance leads to increase the PDR of the network. This adaptive operation of the proposed SI routing protocol increased network lifetime significantly.

Figure 7 depicts the average simulation time “end-to-end delay” in each round for the two topological areas A1 and A2. From these figure, simulation time for proposed swarm routing is approximately equal or little less than of Dijkstra routing. So that applying the proposed swarm routing will not affect on computation time for finding the optimal path for sending packets from source to destination.

TABLE I. NETWORK LIFETIME, PARTITION TIME, AND PDR

Routing Technique	Area	Lifetime	Partition Time	PDR
Dijkstra Routing	A1	1001	2452	1
Proposed SI without AIS	A1	2650	More than 10000	0.877
Proposed SI Routing	A1	2259	2260	1
Dijkstra Routing	A2	602	1498	1
Proposed SI without AIS	A2	1828	More than 10000	0.881
Proposed SI Routing	A2	1789	1792	0.998

### VI. CONCLUSION

WSNs available with limited source power through their life cycle. Since the battery of the sensor node cannot be replaced or recharged, energy preservation occupies first crucial problem in designing WSN infrastructure. This paper presents an adaptive routing technique for maximizing WSNs lifetime using ACO algorithm and the AIS. The AIS is used for solving packet LOOP problem and to control route direction, so that the packet will not sent far away from the sink. While ant colony algorithm in the suggested routing algorithm has used for determining optimum routes for sending data packets.

The proposed energy efficient routing protocol ensures that optimum paths from nodes towards base station is determined along with energy balance. Two topological areas have been used for simulation, which are topological areas A1 and A2. Simulations demonstrate that suggested energy efficient routing technique has better performance against the Dijkstra routing. Simulation results show an increase in network lifetime of about 1.2567 for area A1 and 1.9717 for area A2. Simulation results show that without using artificial immune system for route direction control, the PDR is decreased to about 0.877 for area A1 and 0.881 for area A2. This prove the importance of introducing artificial immune system for route direction control. Our experiments showed that without the use of the artificial immune route direction control, the goal of the routing protocol would not be satisfied. Simulation results show that the lifetime is maximized and the energy exhaustion is balanced.

The efficiency of suggested method depicts good contribution in field of maximizing network lifetime using

adaptive routing techniques. Simulation results prove the generality of the proposed energy efficient routing technique, so that the proposed routing technique could be used for any design framework.

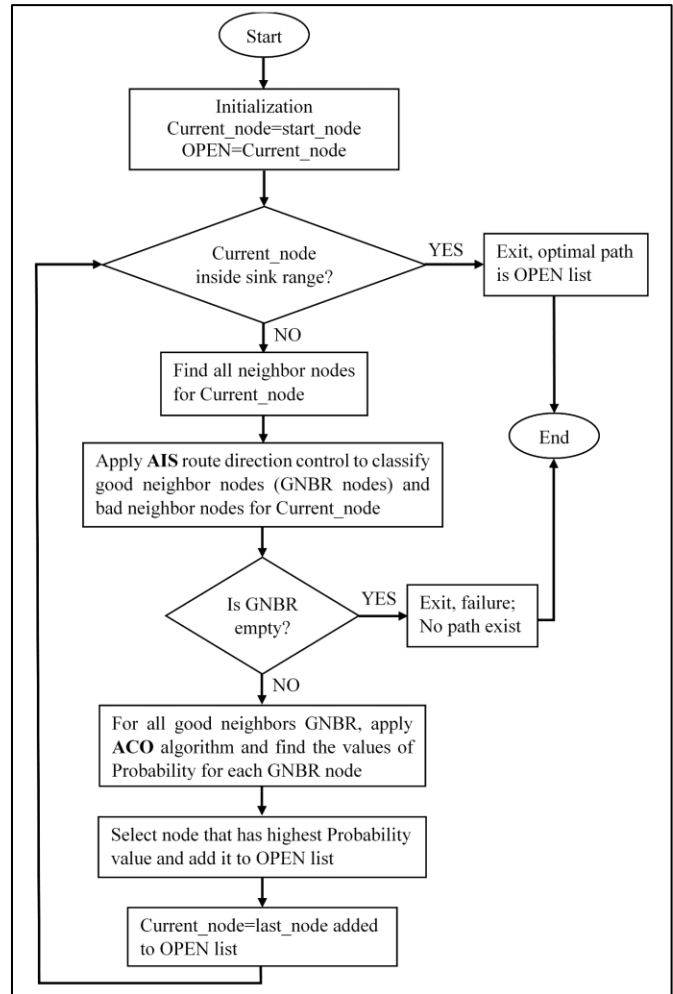


Fig. 1. General structure of the proposed method

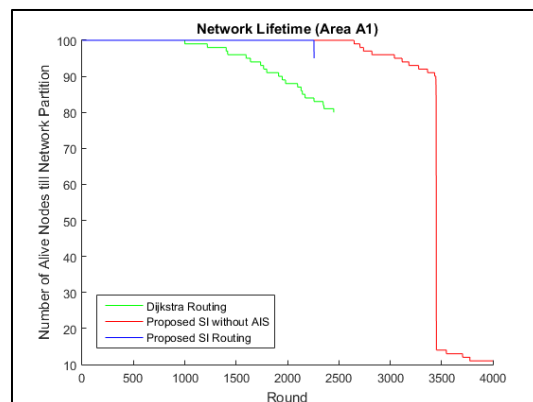


Fig. 2. (a). Number of Alive Nodes for Areas A1

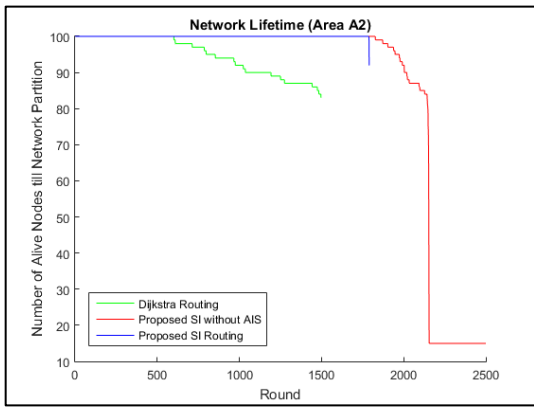


Fig. 2. (b). Number of Alive Nodes for Areas A2

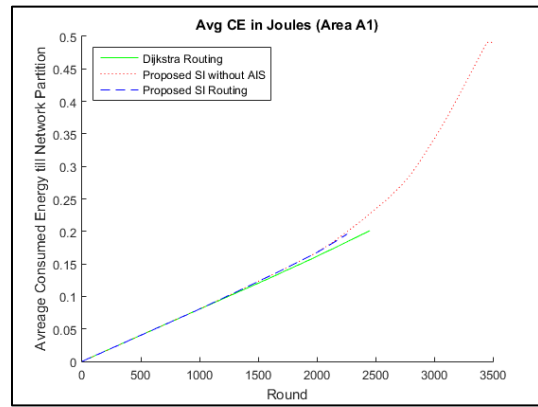


Fig. 4. (a). Average Consumed Energy for Areas A1

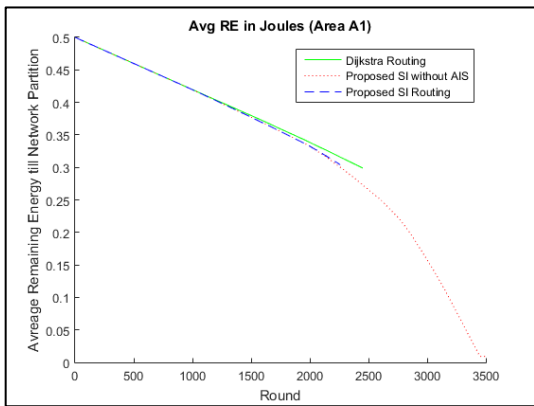


Fig. 3. (a). Average Remaining Energy for Areas A1

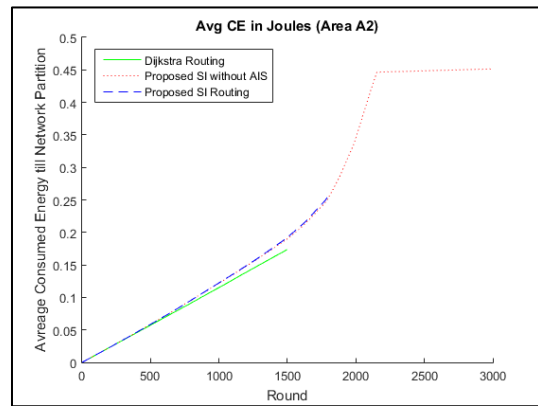


Fig. 4. (b). Average Consumed Energy for Areas A2

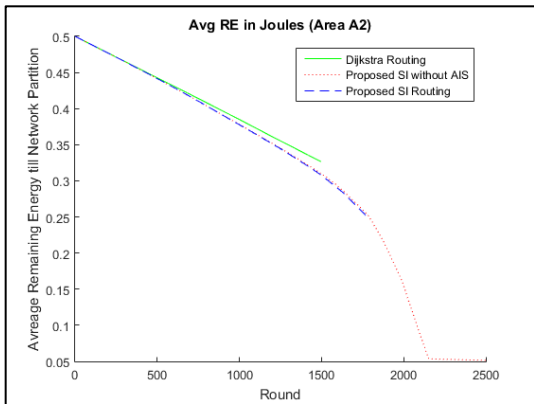


Fig. 3. (b). Average Remaining Energy for Areas A2

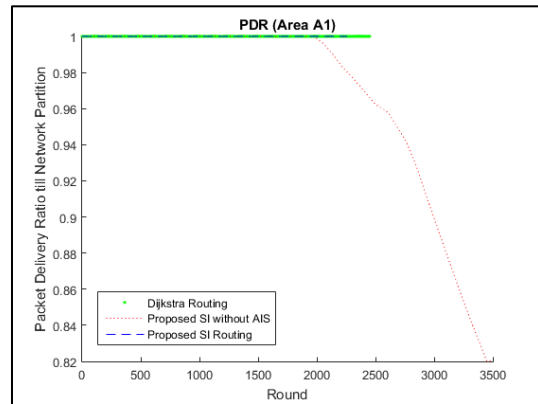


Fig. 5. (a). Packet Delivery Ratio for Areas A1

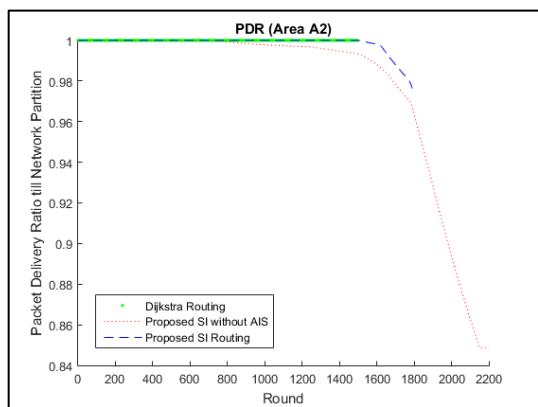


Fig. 5. (b). Packet Delivery Ratio for Areas A2

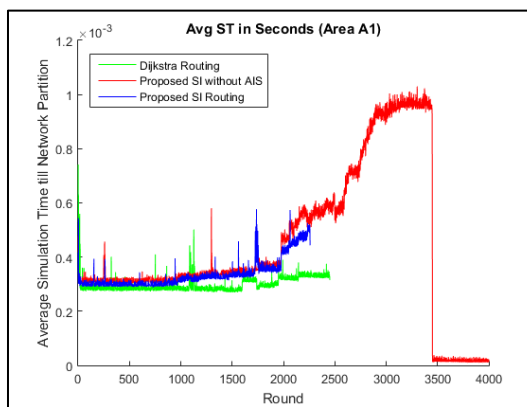


Fig. 7. (a). Average Simulation Time for Areas A1

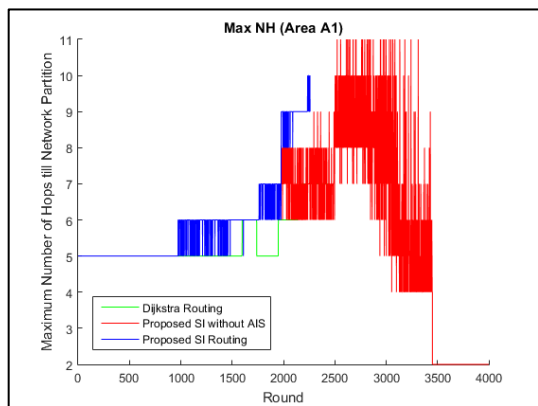


Fig. 6. (a). Maximum Number of Hops for Areas A1

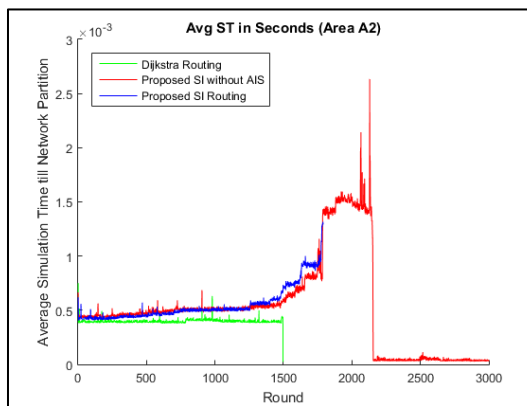


Fig. 7. (b). Average Simulation Time for Areas A2

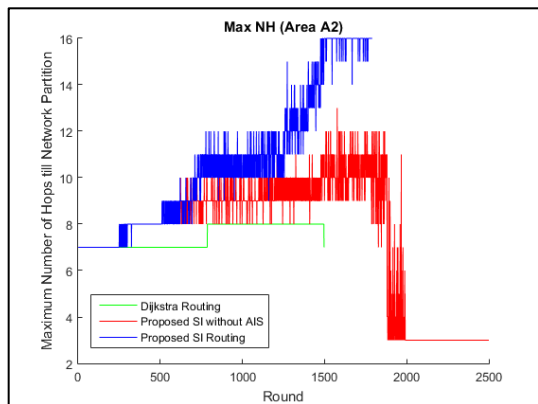


Fig. 6. (b). Maximum Number of Hops for Areas A2

REFERENCES

- [1] Ian F. Akyildiz and Mehmet Can Vuran, "Wireless Sensor Network", John Wiley & Sons Ltd, 2010.
- [2] A. Hac, "Wireless Sensor Network Designs", John Wiley & Sons Ltd, 2003.
- [3] Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A Survey on Sensor Networks", IEEE Communications Mag, Vol. 40, No. 8, 2002, pp.102-114.
- [4] Al-Karaki, J.N. Kamal and A.E, "Routing Techniques in Wireless Sensor Networks: A Survey", IEEE Wireless Communication, Vol. 11, 2004, pp.6-28.
- [5] Ananthram Swami, Qing Zhao, Yao-Win Hong and Lang Tong, "Wireless Sensor Networks: Signal Processing and Communications Perspectives", John Wiley & Sons Ltd, 2007.
- [6] Amiya Nayak and Ivan Stojmenovic, "Wireless Sensor and Actuator Networks: Algorithms and Protocols for Scalable Coordination and Data Communication", John Wiley & Sons, Inc, 2010.

- [7] C. Hua and T. P. Yum, "Optimal Routing And Data Aggregation For Maximizing Lifetime Of Wireless Sensor Networks", IEEE ACM Transaction on Network, Vol. 16, No. 4, 2008, pp.892-903.
- [8] H. R. Karkvandi, E. Pecht, and O. Yadid, "Effective Lifetime-Aware Routing In Wireless Sensor Networks", IEEE Sensors Journal, Vol. 11, No. 12, 2011, pp.3359-3367.
- [9] Aniket A. Gurav and Manisha J. Nene, "Multiple Optimal Path Identification Using Ant Colony Optimisation In Wireless Sensor Network", International Journal of Wireless & Mobile Networks, Vol. 5, No. 5, 2013.
- [10] Fatemeh Rismanian Yazdi, "Ant Colony With Colored Pheromones Routing For Multi Objectives Quality Of Services In WSNs", International Journal of Research in Computer Science, Vol. 3, Issue 1, 2013, pp.1-9.
- [11] Ahmed M. Shamsan Saleh, Borhanuddin Mohd Ali, Mohd Fadlee A. Rasid and Alyani Ismail, "A Self-Optimizing Scheme for Energy Balanced Routing in Wireless Sensor Networks Using Sensor Ant", Sensors Journal, Vol. 12, 2012, pp.11307-11333.
- [12] Bharathi M.A and B. P. Vijaya Kumar, "Reverse Game Theory Approach for Aggregator Nodes Selection with Ant Colony Optimization Based Routing in Wireless Sensor Network", International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, 2012, pp.292-298.
- [13] Lianyu Wang, Qinglin Sun, and Hongwen Mal, "Energy Consumption Optimize Based On Ant Colony Algorithm For Wireless Sensor Networks", IEEE 2nd International Asia Conference on Informatics in Control, Automation and Robotics, Vol. 1, pp.17-21, March 6-7, 2010, Wuhan.
- [14] Wenjing Guo, Wei Zhang, and Gang Lu, "A Comprehensive Routing Protocol in Wireless Sensor Network Based on Ant Colony Algorithm", IEEE Second International Conference on Networks Security, Wireless Communications and Trusted Computing, Vol. 1, pp.41-44, April 24-25, 2010, Wuhan, Hubei.
- [15] Wang Guifeng, Wang Yong, and Tao Xiaoling, "An Ant Colony Clustering Routing Algorithm for Wireless Sensor Networks", IEEE Third International Conference on Genetic and Evolutionary Computing, pp.670-673, October 14-17, 2009, Guilin.
- [16] Zhen-wei Shen; Yi-hua Zhu, Xian-zhong Tian, and Yi-ping Tang, "An Ant Colony System Based Energy Prediction Routing Algorithms for Wireless Sensor Networks", IEEE 4th international conference on wireless communications, networking and mobile computing, pp.1-4, October 12-14, 2008, Dalian.
- [17] Yaofeng Wen, Yuquan Chen, and Dahong Qian, "An Ant-based approach to Power-Efficient Algorithm for Wireless Sensor Networks", Proceedings of the World Congress on Engineering 2007, Vol. II, pp.1546-1550, July 2 - 4, 2007, London, U.K.
- [18] Selcuk Okdem and Dervis Karaboga, "Routing in Wireless Sensor Networks Using Ant Colony Optimization", Proceedings of the IEEE First NASA/ESA Conference on Adaptive Hardware and Systems, pp.401-404, June 15-18, 2006, Istanbul.
- [19] Dipankar Dasgupta, "Artificial Immune Systems and Their Applications", Springer-Verlag Berlin Heidelberg, 1999.
- [20] Dipankar Dasgupta, "Advances in Artificial Immune System", IEEE Computational Intelligence Magazine, 2006, pp.40-49.
- [21] S. Forrest, A.S. Perelson, L. Allen, and R. Cherukuri, "Self-Nonself Discrimination in a Computer", Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy, 1994, Los Alamitos, CA: IEEE Computer Society Press.
- [22] Marco Dorigo and Thomas Stützle, "Ant Colony Optimization", Massachusetts Institute of Technology, 2004.
- [23] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy Efficient Communication Protocol For Wireless Microsensor Networks", Proceedings of the 33rd Annually Hawaii International Conference on Systems Sciences, 2000, pp.1-10.

# Improving DNA Computing Using Evolutionary Techniques

Godar J. Ibrahim  
Software Engineering  
College of Engineering  
Salahadin university-Erbil  
Hawler, Kurdistan

Tarik A. Rashid  
Software Engineering  
College of Engineering  
Salahadin university-Erbil  
Hawler, Kurdistan

Ahmed T. Sadiq  
Software Engineering  
College of Engineering  
Salahadin university-Erbil  
Hawler, Kurdistan

**Abstract**—the field of DNA Computing has attracted many biologists and computer scientists as it has a biological interface, small size and substantial parallelism. DNA computing depends on DNA molecules' biochemical reactions which they can randomly anneal and they might accidentally cause improper or unattractive computations. This will inspire opportunities to use evolutionary computation via DNA. Evolutionary Computation emphasizes on probabilistic search and optimization methods which are mimicking the organic evolution models. The research work aims at offering a simulated evolutionary DNA computing model which incorporates DNA computing with an evolutionary algorithm. This evolutionary approach provides the likelihood for increasing dimensionality through replacing the typical filtering method by an evolutionary one. Thus, via iteratively increasing and recombination a population of strands, eliminating incorrect solutions from the population, and choosing the best solutions via gel electrophoresis, an optimal or near-optimal solution can be evolved rather than extracted from the initial population.

**Keywords**—Parallel Computation; DNA Computation Algorithm; Evolutionary DNA Computing Algorithm

## I. INTRODUCTION

Leonard Adleman was the first person to demonstrate that DNA computing could be utilized for computing in a laboratory environment and without using conventional computing devices. As a paradigm for his novel approach, he selected the Hamiltonian Path Problem (HPP) to obtain solutions via experimental DNA tests [1]. It is of interest to mention that further or more rapid progress would have been achieved if he had selected an easier problem. Subsequently, he selected the HPP on the use of conventional computers [2]. This choice opened stimulating avenues and allowed other researchers to think more favorably of DNA computing. In doing so, he considered that the power of this technique was great and involved substantial prospects for parallel computing, as is feasible via operations with DNA. Adelman is now considered the founding father of DNA computing [3].

Currently, DNA computing is an interdisciplinary area in which ecologists, biologists, computer scientists, physical scientists, mathematicians, chemists, and other related specialists identify interesting problems that may be useful for the theoretical and practical sides of DNA computing [4]. DNA computers are essentially assortments of chosen DNA strands. The combinations of these strands imply solutions to a given problem that is to be solved. Tools are currently available to

select the preliminary components and winnow candidate solutions.

The potential of parallel processing algorithms is substantial. DNA computing on large problems can involve parallel processing, given a preliminary arrangement and ample DNA. These challenging tasks are easily and effectively accomplished using DNA computing. In contrast, standard computers would require substantial parallelism and more hardware [2]. On the other hand, DNA computing is determined by DNA molecules' biochemical reactions, which they can unsystematically anneal and they might yield unsuitable computations. Therefore, the use evolutionary computation via DNA might be a good solution. Evolutionary computation emphasizes on optimization method and a search which is probabilistic. It mimics the organic evolution models via using operators such as recombination and randomized mutation or so called the progression of interaction between the formation of fresh generations and their assessments and the choice wherever a sole individual in a populace is affected by the surroundings and also by other individuals. It is considered that individuals that can perform better in such circumstances, they will have greater chances for surviving [5, 6].

The contribution of this work is to propose an evolutionary DNA algorithm based on the standard DNA algorithm as it is presented in this paper to solve the shortest path to increase the possibility of having an optimum solution and to improve the average cost of the final paths. In the next section, some of the techniques and approaches that have been used in the area of DNA computing are reviewed and outlined.

## II. RELATED WORKS

DNA computing was primarily conceived by Leonard Adleman; in 1994, Adleman developed the idea of using DNA computing to address the HPP [1]. From the primary investigations and early tests by Adleman, innovations and progress emerged, e.g., a number of Turing machine devices were systematically developed and tested. Although the primary efforts to exploit this fresh methodology unveiled computationally complex problems, it was rapidly and conclusively shown that the methods were clearly inapplicable to this type of computational algorithm [3, 4, 7, 8].

In 1997, a research team led by Ogihara and Ray recommended the analysis of Boolean circuits, it is clear that Nand Boolean circuits cover only Nand gates. The research

team showed that the relationship between the logarithm of the maximum fan-out of the Boolean circuit and the runtime slow-down is proportionate, also they showed that the relationship among the product of the size and the maximum fan-out and the space complexity is proportionate [ 9, 10, 11, 7, 12, 13].

In 2002, researchers from Israel announced a computing device based on molecular programming. This computing device was constructed from DNA molecules and enzymes in contrast to the silicon materials of microchips and integrated circuits [14, 12]. Shapiro and his research team wrote in Nature that they had created a deoxyribonucleic acid computer [14]. The device was able to locate tumor-related entities within a cell, and the team was able to produce drugs that conferred cancer resistance whenever the disease was detected [14, 15].

Another study was conducted to determine SSCP sensitivity in detecting factor IX mutation [16]. The study investigated the blood of hemophilia B patients in Iran. Phenol, chloroform and other reagents were used to extract DNA. The gene regions were amplified using primer pairs and PCR. PCR fragments with improved flexibility were obtained. The study concluded that SSCP sensitivity was high in the system investigated [16].

In 2006, Dimitrova outlined and summarized examples for various DNA computing applications such as aqueous computing, molecular computing, DNA Turing machines, and the nascent field of synthetic biology [15]. In 2008, Abdulla successfully inserted a heuristic search in a DNA computing algorithm to generate better efficiency and flexibility. That study improved the DNA search technique by increasing the number of solutions and reducing the running time and memory capacity. The modifications of A\* and alpha beta using DNA produced better results than ordinary A\* and alpha beta algorithms [17].

Hari and his research team have suggested an advanced and efficient technique for addressing the problem of minimum vertex cover using a DNA computing algorithm. They suggested that a DNA computing algorithm could make it possible to address problems that are intractable on silicon computers. Nevertheless, the study stated that DNA computing algorithms are not feasible for solving simple problems due to their high degree of parallelism. Thus, computer scientists are required to design and elaborate more DNA computing algorithms [18]. The same research team has suggested that appropriate enzymes and legitimate approaches will dynamically shape biology to address more subtle problems and reduce the amount of error associated with the use of the algorithm.

The focus of this research paper is to incorporate an evolution strategy into DNA computing based only on the crossover operator and strategy parameter. This approach is expected to enhance or optimize the quality of the final solutions by increasing the size of the final solutions and also by evolving the most correct solutions to obtain an optimum or near optimum solution(s). This optimization could take place regardless of the increase in the time and memory capacity of the modified algorithm.

The structure of this paper is as follows. In the next section, DNA computing algorithms and operations are described. Then, a DNA interpretation for the problem of the shortest path is described in detail. Then after, the proposed evolutionary DNA computing method is established. Next, experimental results are described in detail. Finally, the conclusions are outlined. A glossary or list of terminology is presented in appendix A at the end of this paper.

### III. DNA COMPUTING ALGORITHM AND OPERATIONS

Bioinformatics is now viewed as the study of information stored in DNA. The strings of letters correspond to combinations of the four bases A (adenine), T (thymine), G (guanine) and C (cytosine). These strings transmute information via convolution operations on the unit cell [18, 15]. DNA polymerase is the key enzyme. With a specified strand of DNA under suitable circumstances, this enzyme generates the complementary strand, another Watson-Crick DNA sequence in which the letter C stands opposite G, G stands opposite C, A stands opposite T and T stands opposite A. From the molecular sequence CATGTC, for example, the new molecular sequence GTACAG is created by DNA polymerase. DNA is regenerated by DNA polymerase. This capability allows cells to regenerate and eventually permits the investigator to make a replica of the original object of study or analysis. The replication of DNA via DNA polymerase is the most important life process. DNA polymerase is, in essence, a nano-machine. It links to DNA strands, it reads each base, and, as it passes, it writes the complementary information as a fresh, lengthening strand of DNA [18, 15].

The resemblance between the Turing machine and DNA polymerase is striking. It is known that Alan Turing created and designed a computer in the shape of a toy, which was later called the Turing machine. Originally, the device was intended to be conceptual and appropriate for precise mathematical examination.

Thus, it was meant to be truly simple, and Alan Turing fully succeeded. In one account, the Turing machine is described as a limited control device consisting of pairs of tapes. The limited control device moves sideways through the tape input, reading information. It moves sideways through the output of the tape, which reads and writes other data. It is noted that the limited control device is programmed with basic instructions; it would be easy to code programs for reading the letter strings (A, T, C, G) on the input of the tape and produce the complementary Watson-Crick string as the tape output. Thus, a Turing device can be used for coding different programs, e.g., to produce the complementary Watson-Crick strings or to play Chess. The key operations of DNA computing used in a DNA algorithm are defined below [19, 20, 17, 21, 15, 22, 23, 24, 25]:-

*a) Watson-Crick pairing, as mentioned above, is pervasive; it is obvious that any DNA strand has its Watson-Crick complement. The pairing will occur when a DNA molecule encounters the complement of the original Watson-Crick strand. Later, both DNA strands will be annealed. Both strands join to produce the well-known double helix.*

b) Polymerases allow the copying of information from one molecule to another. The DNA polymerase is able to generate strands of DNA complementary to the original Watson-Crick DNA string.

c) Ligases connect molecules. This concept can be illustrated with DNA ligase, which creates one single strand from two DNA strands. DNA ligase can be utilized in the cell to repair disruptions or breaks that can occur in strands of DNA. This phenomenon can be observed when skin cells are exposed to certain types of light.

d) Nucleases break down and thereby repress deoxyribonucleic and ribonucleic acids.

e) Gel electrophoresis serves to analyze and segregate DNA, RNA and protein macromolecules. It allows heterogeneous DNA molecules to be run on a gel with an electric current and identified.

f) Synthesis. Sequences of DNA can be written on paper and sent to a synthesis facility. The synthesis facility then returns a tube containing 1018 DNA molecules within a few days. Each of the molecules contains the specific sequence requested. Sequences of approximately 100 nucleotides in length can be synthesized dependably in the laboratory. The DNA is delivered in dry form in a narrow tube, and the molecules appear white in color.

#### IV. DNA SOLUTION FOR THE SAMPLE OF NETWORK DIAGRAM

Let a directed graph be given by  $Graph = (Ver, Edg)$ , where  $Ver$  is defined as a group of vertices and  $Edg$  is defined as a group of weights (See Fig. 1). A sequence of these vertices such as  $Ver1, Ver2, Ver3... Vern$  can define a basic path, as long as  $Ver1, Ver2, Ver3... Vern$  are considered distinct [2, 8, 25, 16]. The path length can obviously be calculated by adding up every edge weight in the path. Of course, considering walking from a specific source to a specific destination, the shortest simple path is the path that has the least weight among all of the paths in the graph. The objective is to minimize the total cost in the directed path. This is, of course, the simple shortest path going all the way from the key source to the destination. The shortest path problem is a network of several nodes and edges that are used to link all of the nodes, and the problem can be solved using the standard DNA algorithm.

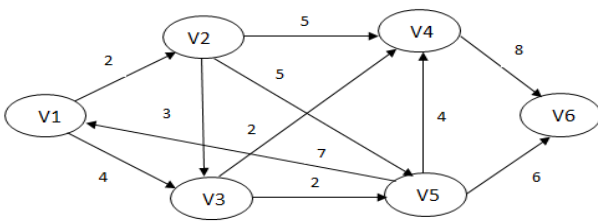


Fig. 1. Sample of Network Diagram

The standard DNA algorithm aims to handle the key DNA operation stages; these stages include the coding of the problem, which is performed in DNA, the production of

random solutions, the amplification of the random DNA solutions using PCR, the elimination of repeated nodes using SSCP, and finally the sorting of the final solutions using electrophoresis (see Fig. 2).

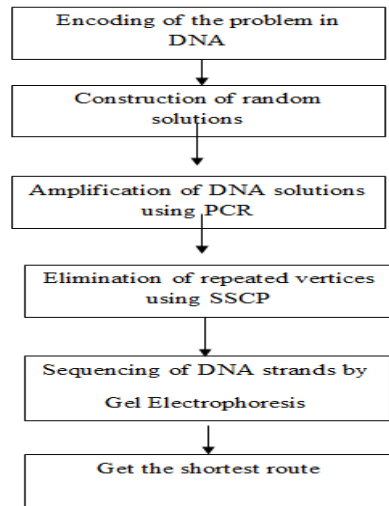


Fig. 2. Shows Block Diagram of Normal DNA Algorithm for Solving Shortest Path Problem

The key stages of the standard DNA algorithm can be described below [2, 16, 15, 26] :-

##### a) DNA problem representation (Coding)

It is obvious that in the graph, each vertex  $Ver_i$  must be linked to a specific palindromic 10-mer DNA sequence, which is represented via  $Ver_i$ . Therefore, every single edge  $Ver_i \rightarrow Ver_j$  (in the directed graph, a complementary oligonucleotide sequence of a 3' 5-mer of  $Ver_i$  is tracked via a complimentary sequence of a 5' 5-mer of  $Ver_j$ ) can be combined.

##### b) Building of random pathways (Construction)

It can be said that every oligonucleotide coding vertex and every oligonucleotide coding edge must be combined to build random paths in the graph; tweaks or a tweak for each of the various sequences can be selected and are kept inside test tubes.

##### c) DNA path intensification using PCR (Amplification)

In this stage, the process of intensification is performed. It would begin with the source of the vertex and would end at the destination of the vertex. Note that 2 specific primers can be designed to target the source of the vertex and destination of the vertex, which will further amplify the PCR response.

##### d) Dismissal of repetitive vertices (Elimination)

SSCP is a simple and common practice of mutation detection. This process prevents nodes from reappearing within the strands of DNA. Basically, the PCR serves to proliferate and aggregate the region of interest. The subsequent DNA would be separated via electrophoresis to produce single-stranded molecules.

##### e) Sequencing of the DNA strands

At this stage, the strands achieved in stage 4 are sequenced. By reading the sequence, the weight of each strand is defined.



The desired solution is considered to be the path that has the least weight.

Applying the normal DNA algorithm to solve the shortest path problem, the above key operational stages can be expressed as follows:

a) DNA problem representation (Coding)

This can be performed via random synthesis of a palindromic 10-base strand of DNA for every single vertex, keeping in mind that *Veri* signifies the *i*th vertex (See Table 1). Thus, to represent every edge, *Veri* → *Verj*, a palindromic 10-base strand of DNA can be synthesized, which is a sequence consisting of the 3' 5-mer complement of *Veri* and the 5' 5-mer sequence complement of *Verj*. The result of the graph (see Fig. 1) is shown in Table 2.

b) Building of random pathways (Construction)

In this stage, all sequences that correspond to both vertices *Veri* and edges *Veri* → *Vj*, as synthesized in stage 1, must be stored in test tubes to allow ligation. Subsequently, ligation will occur when both sequences of the DNA *Ver1* (*GCGAGATCTG*) and DNA Edge *Ver1* → *Ver2* (*TAGACCTTCA*) accidentally come into contact with each other.

The fact that the earlier sequence ended via *ATCTG* and the latter sequence started via *TAGAC* is important because these two sequences are complementary. Thus, the annealing process will occur. If the resultant composite encounters a sequence that matches *Ver2* (*GAAGTCAGTC*), at that point, it may also be able to join the composite, as the earlier sequence complements the later starting sequence. It can be observed that composites can grow lengthwise when edges of DNA sequences are joined together via a vertex of DNA sequences. Accordingly, the paths can be expressed as follows:-

*Ver3* → *Ver4* →  
*GTCACGACGAGCCATAGACC*  
*Ver1* → *Ver2* → *Ver3* → *Ver5* → *Ver4* → *Ver6*  
*GCGAGATCTGGAAGTCAGTCGTCACGACGAGTTCGTTA*  
*GGCCATAGACCGATGAGAGTA*  
*Ver2* → *Ver3* → *Ver5* → *Ver4*  
*GAAGTCAGTCGTCACGACGAGTTCGTTTAGGCCATAGAC*  
*C*  
*Ver1* → *Ver2* → *Ver3* → *Ver5*  
*GCGAGATCTGGAAGTCAGTCGTCACGACGAGTTCGTTA*  
*G*  
*Ver1* → *Ver3* → *Ver4* → *Ver6*  
*GCGAGATCTGGTCACGACGAGCCATAGACCGATGAGAG*  
*TA*  
*Ver1* → *Ver2* → *Ver3* → *Ver5* → *Ver4* → *Ver6*  
*GCGAGATCTGGAAGTCAGTCGTCACGACGAGTTCGTTA*  
*GGCCATAGACCGATGAGAGTA*  
*Ver1* → *Ver2* → *Ver5* → *Ver1* → *Ver3* → *Ver5* → *Ver6*  
*GCGAGATCTGGAAGTCAGTCGTTAGGCGAGATCT*  
*GGTCACGACGAGTTCGTTTAGGATGAGAGTA*  
*Ver1* → *Ver2* → *Ver3* → *Ver5* → *Ver6*  
*GCGAGATCTGGAAGTCAGTCGTCACGACGAGTTCGTTA*  
*GGATGAGAGTA*  
*Ver2* → *Ver3* → *Ver4* → *Ver6*

*GAAGTCAGTCGTCACGACGAGCCATAGACCGATGAGAG*  
*TA*  
*Ver1* → *Ver3* → *Ver5* → *Ver4*  
*GCGAGATCTGGTCACGACGAGTTCGTTTAGGCCATAGA*  
*CC*  
etc.

c) DNA path intensification using PCR (Amplification)

The generation of all paths that can hold both source and destinations (*Vers* and *Verd*) can be accomplished via DNA path intensification using PCR. This process occurs as follows:

- I. If the source *Ver1* and the destination *Ver6* are selected, at that point, both primers that correspond to the source and destination (*Ver1* (*GCGAGATCTG*) and *Ver6* (*GATGAGAGTA*)) are added to the resulting solution. Then, the PCR can occur.
- II. Additionally, the *GCGAGATCTG* primer would anneal to its target sequence that is established in the *Ver1* → *Ver6* path; a similar phenomenon will occur with the other primers.

Consequently, immediately after the accomplishment of PCR, every path from the *Ver1* source to the *Ver6* destination can be expressed as follows:-

*Ver1* → *Ver2* → *Ver3* → *Ver5* → *Ver6* →  
*GCGAGATCTGGAAGTCAGTCGTCACGACGAGTTCGTTA*  
*GGATGAGAGTA*  
*Ver1* → *Ver2* → *Ver3* → *Ver5* → *Ver4* → *Ver6*  
*GCGAGATCTGGAAGTCAGTCGTCACGACGAGTTCGTTA*  
*GGCCATAGACCGATGAGAGTA*  
*Ver1* → *Ver2* → *Ver5* → *Ver1* → *Ver3* → *Ver5* → *Ver6*  
*GCGAGATCTGGAAGTCAGTCGTTAGGCGAGATCT*  
*GGTCACGACGAGTTCGTTTAGGATGAGAGTA*  
*Ver1* → *Ver2* → *Ver4* → *Ver6*  
*GCGAGATCTGGAAGTCAGTCGCCATAGACCGATGAGAG*  
*TA*  
*Ver1* → *Ver3* → *Ver5* → *Ver6*  
*GCGAGATCTGGTCACGACGAGTTCGTTTAGGATGAGAG*  
*TA*  
*Ver1* → *Ver3* → *Ver5* → *Ver4* → *Ver6*  
*GCGAGATCTGGTCACGACGAGTTCGTTTAGGCCATAGA*  
*CCGATGAGAGTA*  
*Ver1* → *Ver2* → *Ver5* → *Ver6*  
*GCGAGATCTGGAAGTCAGTCGTTAGGATGAGAGT*  
*A*

d) Dismissal of repetitive vertices (Elimination)

In this stage, the rule is obviously not to allow the vertices to reappear. In other words, repetition is not allowed; thus, nodes that are repeated for the second time will be dismissed.

The process of elimination can be performed via a single-stranded conformation polymorphism approach (SSCP). A hairpin structure can be formed via the series. It can hold reappearing nodes that can be connected to their corresponding split ends. The strands that hold hairpin loops will eventually be eliminated via the SSCP approach. Thus, the representation for all paths that have vertex repetition can be expressed as follows:-

Ver1 → Ver2 → Ver5 → Ver1 → Ver3 → Ver5 → Ver6  
GCCGAGATCTGGAAGTCAGTCGTTTCGTTTAGGCCGAGATCT  
GGTCACGACGAGTTCGTTTAGGATGAGAGTA

It can be observed that with a path Ver1 → Ver2 → Ver5 → Ver1 → Ver3 → Ver5 → Ver6, the vertices Ver1 and Ver5 are repeated. Thus, subsequently disregarding strands with loops, the achieved paths are expressed as follows:

Ver1 → Ver2 → Ver3 → Ver5 → Ver6  
GCCGAGATCTGGAAGTCAGTCGTCACGACGAGTTCGTTT  
GGATGAGAGTA

Ver1 → Ver2 → Ver3 → Ver5 → Ver4 → Ver6  
GCCGAGATCTGGAAGTCAGTCGTCACGACGAGTTCGTTT  
GGCCATAGACCGATGAGAGTA

Ver1 → Ver2 → Ver4 → Ver6  
GCCGAGATCTGGAAGTCAGTCGCCATAGACCGATGAGAG  
TA

Ver1 → Ver3 → Ver5 → Ver6  
GCCGAGATCTGGTCACGACGAGTTCGTTTAGGATGAGAG  
TA

Ver1 → Ver3 → Ver5 → Ver4 → Ver6  
GCCGAGATCTGGTCACGACGAGTTCGTTTAGGCCATAGA  
CCGATGAGAGTA

Ver1 → Ver2 → Ver5 → Ver6  
GCCGAGATCTGGAAGTCAGTCGTTTCGTTTAGGATGAGAGT  
A

e) Sequencing of the DNA strands. In this stage, each weight path achieved in stage 4 is defined by interpreting the equivalent sequence.

Ver1 → Ver2 → Ver4 → Ver6

Overall weight for path is 15

Ver1 → Ver2 → Ver3 → Ver5 → Ver6

Overall weight for path is 13

Ver1 → Ver2 → Ver3 → Ver5 → Ver4 → Ver6

Overall weight for path is 19

Ver1 → Ver3 → Ver5 → Ver6

Overall weight for path is 12

Ver1 → Ver3 → Ver5 → Ver4 → Ver6

Overall weight for path is 18

Ver1 → Ver2 → Ver5 → Ver6

Overall weight for path is 13

Note that Ver1 → Ver3 → Ver5 → Ver6 path produces 12; this is the least weight cost. Therefore, Ver1 → Ver3 → Ver5 → Ver6 is our solution to the problem path, which holds an equivalent DNA sequence that can be described as follows:

GCCGAGATCTGGTCACGACGAGTTCGTTTAGGATGAGAG  
TA

## V. PROPOSED EVOLUTIONARY DNA COMPUTING

The Java programming language is used to design a simulation of DNA computing to solve the problem of the simple shortest path. The standard DNA algorithm is implemented; then, the cleaning stage is used to obtain the desired solutions randomly. However, the produced solutions that are not desirable are discarded through this stage. The shortest path is established; however, there are some restrictions that must be addressed:-

1) It is clear that the DNA algorithm generates random solutions that are governed by chance with respect to the DNA strands meeting each other or not; thus, the DNA algorithm might not produce all of the potential solutions as there is no evolutionary progress involved here to produce solutions via progress within considerable sequence populations arising from the DNA.

2) In case the DNA algorithm cannot produce all of the potential solutions, identification of the best solution is not assured.

3) The random generation solution size can be amplified; thus, the number of potential solutions could also be amplified. This process will allow us to obtain more final solutions; however, the downsides are that the search process will take more time and there is a greater need for memory capacity. Still, the best solution is not assured.

There must be other ways to enhance the standard normal DNA computing algorithm to obtain more diversity in the scope of the produced solutions, to create more correct solutions and also to obtain the best solution. In this regard, a method is suggested to conglomerate DNA computing with an evolutionary algorithm (Evolutionary DNA Computing). Evolutionary algorithm features are used in this research to produce solutions through progress within substantial sequence populations arising from the DNA. This evolutionary algorithm will increase the dimensionality of the system by replacing the customary filtering approach with an evolutionary approach. Thus, the best solutions might be obtained through iterative intensification, recombining strand populations, eliminating inappropriate solutions included in the population, and selecting the best solutions through gel electrophoresis instead of mining them from the preliminary population.

This proposed improvement of the algorithm has four modifying operations (See Fig. 3). Each operation has an effect on the algorithm, while they all share the same representation of the knowledge. The four operations are the following:-

1) Adding/Replacing the start/end of the PCR-dropped strands

The normal DNA algorithm is modified by adding/replacing the start/end of the PCR-dropped strands. The first level of enhancement is finished by obtaining all of the dropped PCR solutions, adding a start node to the beginning of the solution strand, and adding an end node to the end of the solution strand. By performing this operation, the PCR-dropped solution can address PCR solutions in such a manner that the chance of obtaining more final solutions will increase. Another level that is accomplished using the same function is to switch the start and end nodes of the dropped PCR solution strand by the desired start/end nodes. This step can add diversity to the solution space by increasing the number of PCR solutions and consequently increasing the chance of obtaining many solutions in the end. These 2 modifications are solely applicable to the Shortest Path problem. Details of the pseudo code snippet are shown below.

**Replace and Add the start/end of the PCR solution:**

**Input: Dropped PCR solution**

**Output: Added PCR Test Tube Solution**

```

For each dnaStrand in the PCR Test Tube
  If dnaStrand.length > Size
    newStrand= replaceStartEnd (dnaStrand)
    Add newStrand to PCR Test tube
    newStrand= addStartEnd (dnaStrand)
    Add newStrand to PCR Test tube
  End For
replaceStartEnd (dnaStrand)
replacedStrand= dnaStrand.substring(Size, dnaStrand.length()-Size)
  Return startNodeStrand+replacedStrand+endNodeStrand
addStartEnd (dnaStrand)
  Return startNodeStrand+ dnaStrand +endNodeStrand
replacedStrand= dnaStrand.substring(Size, dnaStrand.length()-Size)
  Return startNodeStrand+replacedStrand+endNodeStrand
addStartEnd (dnaStrand)
  Return startNodeStrand+ dnaStrand +endNodeStrand

```

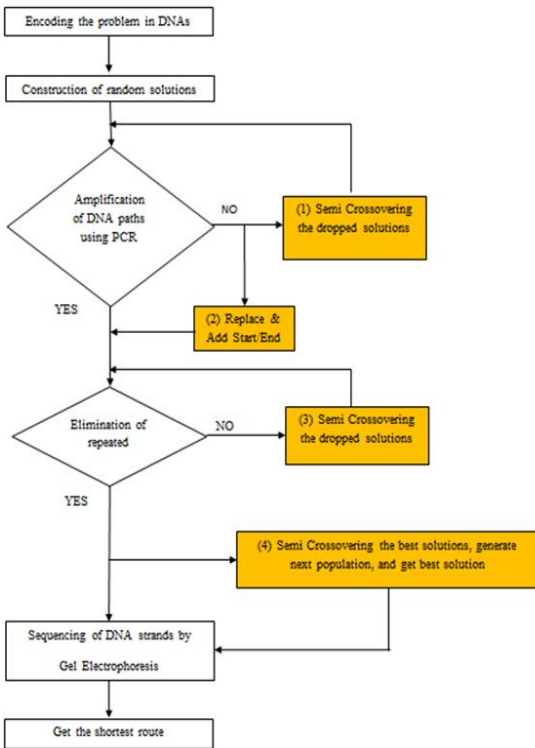


Fig. 3. The proposed technique

### 2) Crossovers Dropped in the PCR Solutions

Modifying the normal DNA algorithm by adding semi-crossovers of the dropped PCR strands involves applying another modification in the algorithm. The process works by obtaining the dropped PCR solutions and randomly selecting two nodes in the solution strand to be replaced by two random nodes obtained from the set of original nodes. This step is similar to a semi-crossover operation for increasing the possibility of obtaining more solutions. The output of this function is sent to the PCR function to obtain the correct PCR solutions and to send them to the next function, which is the SSCP function. The details of the pseudo code snippet are shown below.

#### PCR Semi-Crossover:

**Input:** Dropped PCR Solution TT, Nodes TT  
**Output:** Added PCR Test Tube Solution

```

For each dnaStrand in PCR Dropped Test Tube
  If dnaStrand.length > Size
    While no termination do // two crossovers or break
      Get random nodeS in dnaStrand
      Get random NodeS from initial Test Tube
      Replace nodeS with NodeS
    Od
    If isPCR (dnaStrand) // isPCR function already available
      Add dnaStrand to PCR Test Tube
    End If
  End If
End For

```

### 3) Crossovers Dropped in SSCP Solutions

Modifying the normal DNA algorithm with semi crossovers of the dropped SSCP strands proceeds as follows:-

This modification is the same as the PCR Semi-Crossover but is applied on dropped SSCP solutions, and the difference is that the two random nodes are semi-crossed over, but not the start/end node, because there are already correct nodes in the solutions strand. The output of this function is sent to the SSCP function to obtain the correct SSCP solution from within the set. Details of the pseudo code snippet are shown below.

#### SSCP Semi-Crossover:

**Input:** Dropped SSCP Solution TT, Nodes TT

**Output:** Added SSCP Test Tube Solution

```

For each dnaStrand in the SSCP Dropped Test Tube
  If dnaStrand.length > Size
    While no termination do // two crossovers or break
      Get random nodeS in dnaStrand but not start/end node
      Get random NodeS from initial Test Tube
      Replace nodeS with NodeS
    Od
    If isSSCP (dnaStrand) // isSSCP function already available
      Add dnaStrand to SSCP Test Tube
    End If
  End For

```

### 4) Evolutionary SSCP

Modifying the normal DNA algorithm with the evolutionary approach:

The following is the real evolutionary improvement to the algorithm: the best SSCP solution and 10% of the other SSCP solutions (not the best ones) are taken from the SSCP solution list. A new generation is made from these selected solutions by crossing over one node of the solution randomly and checking if any solution is generated with a lower cost; it is used if no better solution is generated. The strategy parameter is tuned, and instead of crossing over only one node, two nodes are crossed over, and the evolution of the algorithm is evaluated. The generation of a new population from an initial population is continued until either a better solution is obtained or the termination criterion is met. Details of the pseudo code snippet are shown below.

#### SSCP Evolution Strategy:

**Input:** SSCP Solution Test Tube, Nodes Test Tube

**Output:** Added SSCP Test Tube Solution

```

While no termination do // two crossovers or break
  Get best solution and 10% random solutions
  For i=0 to EdgesTT.length // number of generations
    If (new generation worse than previous one)
      Set crossinOverNodes=2 // strategy parameter setting

```

```
Else  
  Set crossinOverNodes=1  
  makeNewGeneration()  
  Get best solution and 10% random solutions from new generation  
End For  
Get best solution from new generation  
End While  
makeNewGeneration()  
While no termination do  
  makeSemiCrossOver(currentStrand,Nodes TT) // function is already  
  //defined  
  add crossedOver strand to new population  
End While
```

## VI. EXPERIMENTAL RESULTS

It is clear that the evolutionary SSCP improves the algorithm by evolving the solution into a better solution. In the next first, second or third generation, a better solution is generated by semi-crossovers of the nodes in the current generation; occasionally, this occurs in later generations. A better indication of the evolutionary improvement is that in some iteration, a solution is given even when there is no solution found in the normal algorithm.

An improved DNA computing algorithm for solving complex optimization problems is presented in this research study. The algorithm not only shows whether a solution exists but also provides more possible solutions; hence, the likelihood of obtaining the optimum solution is increased. The proposed algorithm might be extended to solve other optimization problems. This will be shown in the test result tables. To improve the algorithm, the focus is on the generation of more solutions rather than decreasing the running time or memory capacity, as current computers have sufficient CPU speed and memory capacity. The variables to be used are defined in the results of the Shortest Path Problem; all of the results in this section are supported by tables and charts to display the intermediate and final results, with statistical curves that represent the comparison between the standard DNA and improved DNA algorithms. The data tables of the results of the DNA Algorithm and Evolutionary DNA Algorithm for solving the Shortest Path problem are found below. Table 3 shows the solution of the standard DNA algorithm for SPP, which is the basic result to be compared with the results generated by the improved DNA algorithm.

Table 4 shows the first improved solution of SPP by the DNA algorithm. As explained in the previous sections, the algorithm is improved by working on the dropped solution at the PCR operation; the first step's desired start node will be added to the beginning of the strand, and the end node is added to the end of the strand. At the second step, the beginning node and end node of the strand will be replaced by the desired start and end nodes. It is clear in the table and by comparing with the previous table that the number of PCR solutions is dramatically increased (which increases the SSCP solutions automatically), which results in having more final solutions and a better average cost for the final paths.

By embedding the crossover operation in the PCR operation, as observed below the table, the number of PCR and SSCP is increased, which again results in an increased number

of final solutions and the improvement of the average cost of the final paths, but the percentage of improvement is less than that of the previous modification. By embedding the SSCP crossover in the SSCP operation, the percent increase in the SSCP solution is even lower; hence, it has a smaller number of final solutions and therefore does not have good improvement in the average cost of the final paths. Until now, replacing Start/End in the PCR operation generates better results than the PCR and SSCP crossover operations. Even combining the PCR and SSCP does not yield good results (Tables 5, 6 and 7).

As is shown in Tables 8, 9 and 10, by Replacing Start/End in the PCR Operation with the PCR and SSCP crossover, the number of final solutions and the average cost of the final paths is improved. Thus, replacing Start/End in the PCR operation improves the standard DNA algorithm much more than the PCR and SSCP crossover. Tables 11 and 12 show the results of the evolutionary SSCP, evolving the resulting SSCP solution through several generations until the best SSCP solution is obtained. Although the number of PCRs, SSCPs and final solutions is low, the average cost of the final route is good; this finding indicates that the performance of the algorithm can be increased by evolving the final solution rather than increasing the search space of the problem. By adding the supportive operations to the evolutionary SSCP operation, the result is improved. There are more solutions in the end; therefore, the possibility of having the optimum or near-optimum solution is increased, and the average cost of the final paths is improved as well (See Tables 13, 14, 15 and 16).

The effect of the modifications on the standard DNA algorithm and the optimization of the algorithm by obtaining more solutions and better results in the end are clearer when the above data tables are converted to many distinct charts. The section below shows the corresponding charts of the main factors of the problems to better highlight the improvement in the algorithm. Below, the figures of the DNA Algorithm and Evolutionary DNA Algorithm for solving the Shortest Path problem can be found. Fig. 4 shows the number of solutions versus the number of nodes for the DNA Algorithm and Evolutionary DNA Algorithm. Clearly, the number of solutions is increased with the proposed evolutionary techniques.

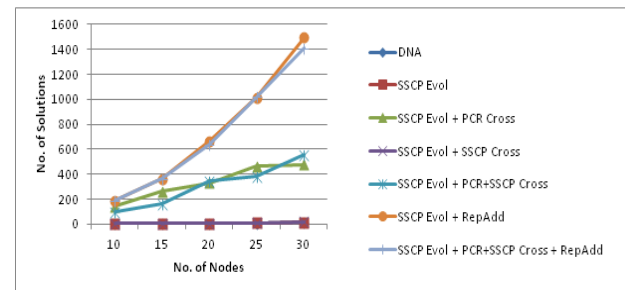


Fig. 4. shows Average No. of PCR Solutions; DNA Algorithm, SSCP Evol, SSCP Evol+PCR Corsss, SSCP Evol+ SSCP Cross, SSCP Evol+PCR+SSCP Cross, SSCP Evol+PCR+SSCP Cross+RepAdd

Fig. 5 shows the number of PCR solutions versus the number of nodes for the DNA Algorithm and Evolutionary DNA Algorithm. The PCRCross+RepAdd produced more solutions than others

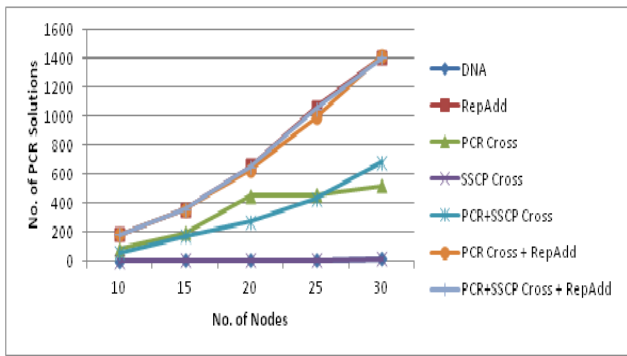


Fig. 5. Average No. of PCR Solutions; DNA Algorithm, RepAdd, PCR Corss, SSCP Cross, +PCR+SSCP Cross, PCR Cross+RepAdd, SSCP Cross+RepAdd

Fig. 6 shows the number of SSCP solutions versus the number of nodes for the DNA Algorithm and Evolutionary DNA Algorithm.

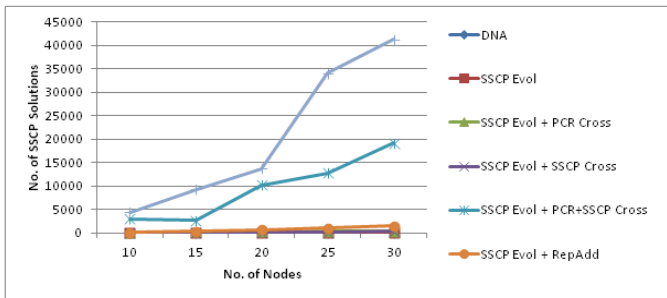


Fig. 6. shows, Average No. of SSCP Solutions; DNA Algorithm, SSCP Evol, SSCP Evol+PCR Corss, SSCP Evol+ SSCP Cross, SSCP Evol+PCR+SSCP Cross, SSCP Evol+PCR+SSCP Cross+RepAdd

Fig. 7 shows the number of SSCP solutions versus the number of nodes for the DNA Algorithm and Evolutionary DNA Algorithm.

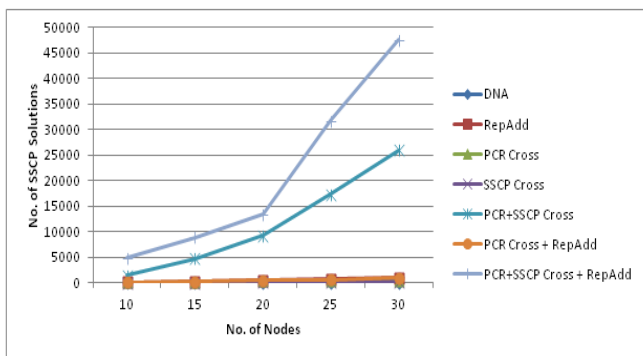


Fig. 7. shows – Average No. of SSCP Solutions; DNA Algorithm, RepAdd, PCR Corss, SSCP Cross, +PCR+SSCP Cross, PCR Cross+RepAdd, SSCP Cross+RepAdd

Fig. 8 shows the number of solutions versus the number of nodes for the DNA Algorithm and Evolutionary DNA Algorithm.

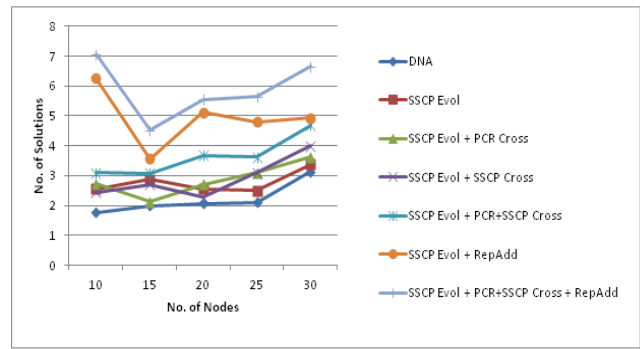


Fig. 8. Average No. of Final Solutions; DNA Algorithm, SSCP Evol, SSCP Evol+PCR Corss, SSCP Evol+ SSCP Cross, SSCP Evol+PCR+SSCP Cross, SSCP Evol+PCR+SSCP Cross+RepAdd

Fig. 9 shows the number of solutions versus the number of nodes for the DNA Algorithm and Evolutionary DNA Algorithm.

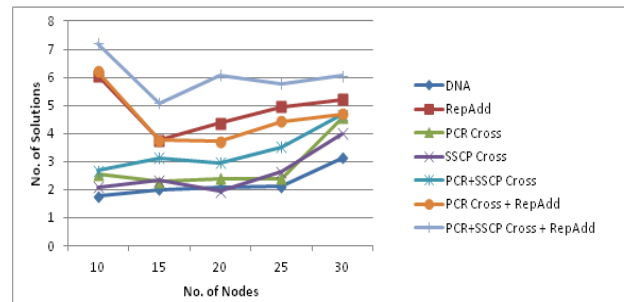


Fig. 9. Average No. of Final Solutions; DNA Algorithm, RepAdd, PCR Corss, SSCP Cross, +PCR+SSCP Cross, PCR Cross+RepAdd, SSCP Cross+RepAdd

Fig. 10 shows the Average path cost versus the number of nodes for the DNA Algorithm and Evolutionary DNA Algorithm. It is clearly seen that the average path cost using DNA algorithm is the most expensive among others.

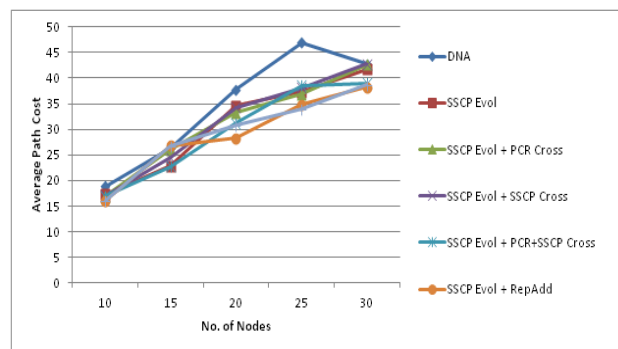


Fig. 10. Average Path Cost; DNA Algorithm, SSCP Evol, SSCP Evol+PCR Corss, SSCP Evol+ SSCP Cross, SSCP Evol+PCR+SSCP Cross, SSCP Evol+PCR+SSCP Cross+RepAdd

Fig. 11 shows the Average Path Cost versus the number of nodes for the DNA Algorithm and Evolutionary DNA Algorithm.

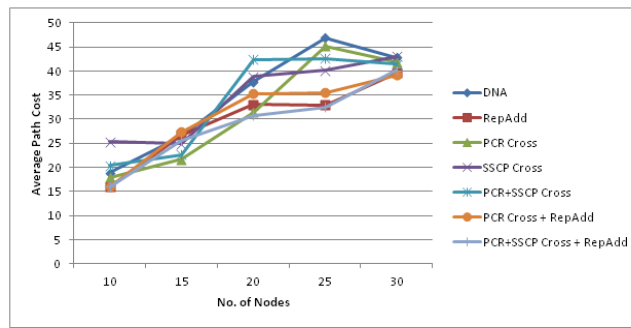


Fig. 11. Average Path Cost; DNA Algorithm, RepAdd, PCR Corss, SSCP Cross, +PCR+SSCP Cross, PCR Cross+RepAdd, SSCP Cross+RepAdd

TABLE I. SHOWS VERTICES SND THEIR COMPLIMENTS

COMPLIMENTS			
Ver1 =	5'GCGAGATCTG3'	Comp: 3'CGCTCTAGAC5'	OR 5'CAGATCTCGC3'
Ver2 =	5'GAAGTCAGTC3'	Comp: 3'CTTCAGTCAG5'	OR 5'GACTGACTTC3'
Ver3 =	5'GTCACGACGA3'	Comp: 3'CAGTGCTGCT5'	OR 5'TCGTCGTGAC3'
Ver4 =	5'GCCATAGACC3'	Comp: 3'CGGTATCTGG5'	OR 5'GGTCTATGGC3'
Ver5 =	5'GTTCTGTTAG3'	Comp: 3'CAAGCAAATC5'	OR 5'CTAAACGAAC3'
Ver6 =	5'GATGAGAGTA3'	Comp: 3'CTACTCTCAT5'	OR 5'TACTCTCATC3'

TABLE II. SHOWS THE REPRESENTATION OF EDGES SND THEIR CORRESPONDING WEIGHTS

Edge	Weight
Ver1 → Ver2 = TAGACCTTCA	2
Ver1 → Ver3 = TAGACCAGTG	4
Ver2 → Ver3 = GTCAGCAGTG	3
Ver2 → Ver4 = GTCAGCGGTA	5
Ver2 → Ver5 = GTCAGCAAGC	5
Ver3 → Ver4 = GACGACGGTA	2
Ver3 → Ver5 = GACGACAAGC	2
Ver4 → Ver6 = CGGTACTACT	8
Ver5 → Ver4 = AAATCCGGTA	4
Ver5 → Ver6 = AAATCCTACT	6

TABLE III. STANDARD DNA ALGORITHM

V	E	LIG	RSG	PCR	SSCP	GELP	AP	SP	RT	MC	drpPCR	pcrCross	pcrGen	drpSSCP	sscpCross	sscpGen	evolSSCPerss
10	18	6248	55	2	2	2	19	13	27	13	0	0	0	0	0	0	0
15	23	26198	100	3	3	2	26	17	46	54	0	0	0	0	0	0	0
20	30	90071	171	4	4	3	38	16	88	172	0	0	0	0	0	0	0
25	39	170626	267	6	5	3	47	14	142	397	0	0	0	0	0	0	0
30	47	338264	386	15	8	3	43	30	290	1077	0	0	0	0	0	0	0

TABLE IV. IMPROVED DNA ALGORITHM USING REPLACING START/END AT PCR OPERATION

V	E	LIG	RSG	PCR	SSCP	GELP	AP	SP	RT	MC	drpPCR	pcrCross	pcrGen	drpSSCP	sscpCross	sscpGen	evolSSCPerss
10	18	5892	55	186	139	28	16	13	53	7	52	0	186	47	0	0	0
15	23	23858	101	360	259	57	27	17	48	48	98	0	360	101	0	0	0
20	30	80279	169	661	498	101	33	16	91	165	164	0	661	163	0	0	0
25	38	164058	266	1064	699	165	33	14	157	712	259	0	1064	365	0	0	0
30	47	348691	381	1409	945	222	40	30	288	1219	367	0	1409	464	0	0	0



TABLE V. IMPROVED DNA ALGORITHM USING PCR CROSS OVER

V	E	LIG	RSG	PCRP	SSCP	GELP	AP	SP	RT	MC	drpPCR	pcrCross	pcrGen	drpSSCP	sscpCross	sscpGen	evolSSCPcross
10	18	6825	55	87	71	5	18	13	46	27	51	4159	84	16	0	0	0
15	23	27003	100	194	150	7	22	17	68	81	95	10268	189	43	0	0	0
20	30	91793	173	454	349	15	31	16	132	610	168	21817	449	104	0	0	0
25	39	173501	267	461	285	18	45	14	227	532	260	37273	455	176	0	0	0
30	47	348751	385	520	300	29	42	30	398	1246	363	54790	499	220	0	0	0

TABLE VI. IMPROVED DNA ALGORITHM USING SSCP CROSS OVER

V	E	LIG	RSG	PCRP	SSCP	GELP	AP	SP	RT	MC	drpPCR	pcrCross	pcrGen	drpSSCP	sscpCross	sscpGen	evolSSCPcross
10	18	5826	55	3	5	2	25	13	25	20	51	0	0	0	45	3	0
15	23	26540	100	3	13	2	25	17	48	212	96	0	0	0	42	11	0
20	30	87468	170	4	47	3	39	16	94	263	165	0	0	1	68	44	0
25	39	168514	264	7	70	3	40	14	148	544	256	0	0	2	248	65	0
30	47	364737	383	16	195	4	43	30	288	1081	367	0	0	5	656	184	0

TABLE VII. IMPROVED DNA ALGORITHM USING PCR+SSCP CROSS OVER

V	E	LIG	RSG	PCRP	SSCP	GELP	AP	SP	RT	MC	drpPCR	pcrCross	pcrGen	drpSSCP	sscpCross	sscpGen	evolSSCPcross
10	18	6098	55	60	1457	5	20	13	47	29	52	4290	58	15	173	1412	0
15	23	27694	99	175	4790	7	23	17	70	96	94	10275	172	51	693	4666	0
20	30	85079	168	272	9312	11	42	16	142	546	162	20665	267	105	2527	9145	0
25	39	166696	265	432	17259	18	43	14	252	621	257	36308	425	192	3903	17019	0
30	47	334954	383	686	26016	37	41	33	458	1309	367	55369	671	310	9978	25640	0

TABLE VIII. IMPROVED DNA ALGORITHM USING PCR CROSS OVER + REPLACING START/END AT PCR OPERATION

V	E	LIG	RSG	PCRP	SSCP	GELP	AP	SP	RT	MC	drpPCR	pcrCross	pcrGen	drpSSCP	sscpCross	sscpGen	evolSSCPcross
10	18	5911	55	187	137	28	16	13	30	19	52	0	187	50	0	0	0
15	23	23973	101	359	257	55	27	17	52	60	97	0	359	102	0	0	0
20	30	82685	169	634	475	98	35	16	95	354	163	0	634	158	0	0	0
25	39	167655	268	995	645	154	36	14	156	513	262	0	995	351	0	0	0
30	48	343511	386	1421	898	205	39	30	288	1008	370	0	1421	523	0	0	0



TABLE IX. IMPROVED DNA ALGORITHM USING SSCP CROSS OVER + REPLACING START/END AT PCR OPERATION

V	E	LIG	RSG	PCR	SSCP	GELP	AP	SP	RT	MC	drpPCR	pcrCross	pcrGen	drpSSCP	sscpCross	sscpGen	evolSSCPcross
10	18	5786	55	186	4793	28	16	13	41	31	52	0	186	50	570	4657	0
15	23	25132	101	360	8347	59	27	17	68	320	98	0	360	94	2554	8081	0
20	30	84607	169	632	14160	100	32	16	120	391	163	0	632	161	4421	13689	0
25	39	164140	266	1010	30632	158	34	14	212	696	259	0	1010	342	7742	29964	0
30	47	346084	380	1397	41136	217	40	30	368	1237	364	0	1397	470	12694	40210	0

TABLE X. IMPROVED DNA ALGORITHM USING PCR+SSCP CROSS OVER + REPLACING START/END AT PCR OPERATION

V	E	LIG	RSG	PCR	SSCP	GELP	AP	SP	RT	MC	drpPCR	pcrCross	pcrGen	drpSSCP	sscpCross	sscpGen	evolSSCPcross
10	18	6523	55	186	4856	29	16	13	71	50	52	0	186	50	429	4719	0
15	23	25262	100	358	8853	57	26	17	120	381	97	0	358	102	3012	8597	0
20	30	82588	170	652	13386	105	31	16	215	745	164	0	652	154	4783	12888	0
25	38	167798	264	1053	31748	162	33	14	369	1209	257	0	1053	360	9255	31055	0
30	47	338903	381	1404	47527	209	40	30	656	2245	366	0	1404	533	12219	46656	0

TABLE XI. IMPROVED DNA ALGORITHM USING EVOLUTIONARY SSCP

V	E	LIG	RSG	PCR	SSCP	GELP	AP	SP	RT	MC	drpPCR	pcrCross	pcrGen	drpSSCP	sscpCross	sscpGen	evolSSCPcross
10	18	6383	55	3	32	3	17	13	1319	1285	51	0	0	0	0	0	11,400
15	23	25712	100	4	37	3	23	17	1721	664	95	0	0	1	0	0	16,847
20	30	88395	171	5	46	3	35	16	2093	876	165	0	0	0	0	0	15,270
25	39	169204	264	7	54	3	37	14	2847	1019	256	0	0	2	0	0	26,411
30	47	340939	387	16	69	4	42	30	4189	855	371	0	0	5	0	0	11,195

TABLE XII. IMPROVED DNA ALGORITHM USING EVOLUTIONARY SSCP + PCR CROSS OVER

V	E	LIG	RSG	PCR	SSCP	GELP	AP	SP	RT	MC	drpPCR	pcrCross	pcrGen	drpSSCP	sscpCross	sscpGen	evolSSCPcross
10	18	6503	55	141	117	5	17	13	1190	1177	51	4416	138	52	0	0	48911
15	23	28113	101	261	223	7	26	17	1560	1208	96	10645	257	71	0	0	18559
20	30	88658	171	328	264	10	33	16	2022	477	165	21173	323	105	0	0	18237
25	39	167883	266	461	326	21	37	14	2619	705	258	36408	454	185	0	0	9635
30	47	340505	385	478	344	26	43	30	3720	795	370	56308	464	193	0	0	33182

TABLE XIII. IMPROVED DNA ALGORITHM USING EVOLUTIONARY SSCP + SSCP CROSS OVER

V	E	LIG	RSG	PCR	SSCP	GELP	AP	SP	RT	MC	drpPCR	perCross	perGen	drpSSCP	sscpCross	sscpGen	evolSSCPcross
10	18	6308	55	2	31	2	17	13	1254	967	52	0	0	0	0	0	17900
15	23	25644	100	3	48	3	24	17	1537	1231	96	0	0	0	48	11	19387
20	30	92930	171	4	76	3	34	16	2130	459	166	0	0	1	36	32	18593
25	39	161309	266	8	125	4	38	14	2712	789	257	0	0	2	316	70	26349
30	47	329616	386	16	369	4	43	30	3500	1120	369	0	0	6	615	301	24955

TABLE XIV. IMPROVED DNA ALGORITHM USING EVOLUTIONARY SSCP + PCR+SSCP CROSS OVER

V	E	LIG	RSG	PCR	SSCP	GELP	AP	SP	RT	MC	drpPCR	perCross	perGen	drpSSCP	sscpCross	sscpGen	evolSSCPcross
10	18	5839	55	97	3086	5	17	17	1324	1150	52	4577	94	32	431	2993	48161
15	23	28748	99	161	2665	7	23	17	1484	1068	96	11258	158	31	1226	2502	22009
20	30	86501	170	345	10233	13	31	16	2086	829	165	21433	340	113	2483	9961	16995
25	39	167654	268	382	12767	17	39	14	2653	1208	261	37471	377	146	3875	12480	9759
30	47	356121	383	555	19193	27	39	30	4100	683	362	54019	536	222	6353	18802	27878

TABLE XV. IMPROVED DNA ALGORITHM USING EVOLUTIONARY SSCP + REPLACE/ADD START/END AT PCR OPERATION

V	E	LIG	RSG	PCR	SSCP	GELP	AP	SP	RT	MC	drpPCR	perCross	perGen	drpSSCP	sscpCross	sscpGen	evolSSCPcross
10	18	6089	55	186	166	29	16	13	1162	1496	52	0	186	49	0	0	24942
15	23	23191	101	361	285	53	27	17	1471	1091	98	0	361	110	0	0	17620
20	30	84220	170	666	541	106	28	16	2003	430	166	0	666	167	0	0	15522
25	39	158626	267	1014	697	152	35	14	2539	890	261	0	1014	367	0	0	28331
30	47	330718	381	1496	1007	211	38	30	2952	1201	365	0	1496	547	0	0	23650

TABLE XVI. IMPROVED DNA ALGORITHM USING EVOLUTIONARY SSCP + PCR+SSCP CROSS OVER + REPLACE/ADD START/END AT PCR OPERATION

V	E	LIG	RSG	PCR	SSCP	GELP	AP	SP	RT	MC	drpPCR	perCross	perGen	drpSSCP	sscpCross	sscpGen	evolSSCPcross
10	18	5696	55	188	4477	30	16	13	1266	1399	52	0	188	48	855	4308	22460
15	23	24291	101	361	9325	58	27	17	1442	1187	98	0	361	99	1484	9029	15070
20	30	81129	169	637	13710	101	31	16	1959	462	165	0	637	163	5782	13195	13177
25	39	153987	267	1018	34081	151	34	14	2542	1165	261	0	1018	383	8935	33396	24580
30	47	347375	385	1410	41275	214	39	30	3079	1194	368	0	1410	473	13073	40279	21016

## VII. CONCLUSIONS

In this paper, the fundamental ideas of DNA computing and evolutionary strategies are presented and elaborated. DNA computing is employed to resolve the shortest path problem. The results of the DNA computing algorithm are obtained and the performance of the algorithm is evaluated. Better results are thereby verified. Thus, a suggested Evolutionary DNA Algorithm was considered to take advantage of the Evolutionary Strategies by being embedded in the normal DNA Algorithm to optimize it and hence obtain better results. The optimization produces better results; this means that the number of solutions is increased; thus, the possibility of obtaining optimum solutions is increased as well. Additionally, because the evolutionary technique is used, the initial resulting solutions are evolved; hence, the average quality of the solution generation after generation is increased.

### REFERENCES

- [1] M. Adleman, "Molecular computation of solutions to combinatorial problems," *Science*, vol. 266, pp. 1021-1024, 1994.
- [2] S. Hari, K. Rajeev and S. Vikas, "An approach towards the solution of NP-Complete Problem," *Report and Opinion*, vol. 3, no. 5, 2011.
- [3] D. Boneh, C. Dunworth, J. Lipton and I. Sgall, "On the computational power of DNA," *Discrete Applied Mathematics*, vol. 71, pp. 79-94, 1996.
- [4] L. Kari, G. Gloor and Y. Sheng, "Using DNA to solve the Bounded Post Correspondence Problem," *Theoretical Computer Science*, vol. 231, no. 2, pp. 192-203, 2000.
- [5] G. Gautam and C. Biswanath, "A cascaded pairwise biomolecular sequence alignment technique using evolutionary algorithm," *Information Sciences*, 2014.
- [6] C. Rudy, M. Buyong, N. Ruth and S. Amarda, "Mapping the Conformation Space of Wildtype and Mutant H-Ras with a Memetic, Cellular, and Multiscale Evolutionary Algorithm," *PLoS Computational Biology*; 2015.
- [7] S. Junnarkar, "In Just a new Drops, A Breakthrough in Computing," *New York Times*; 1997.
- [8] F. Mancini, "New perspectives on the Ising model," *the June*, vol. 45, no. 4, pp. 497-514, 2005.
- [9] M. Ogihara M and A. Ray, "Simulating Boolean circuits on a DNA computer, *Algorithmica*, Published by Springer, vol. 25, pp. 239-250, 1999.
- [10] T. Bäck, N. Kok and G. Rozenberg, "Evolutionary computation as a paradigm for DNA-based computing," In Landweber, L.F. and Winfree, E. (eds), *Evolution as Computation*. DIMACS workshop, Princeton, January 1999. Springer-Verlag, Heidelberg, Germany, pp. 15-40, 2003.
- [11] T. Bäck, "Evolutionary Computation as a Paradigm for DNA-Based Computing," *Natural Computing Series*; 2002.
- [12] S. Lovgren, "Computer Made from DNA and Enzymes," *National Geographic*; 2003.
- [13] H. Ahrabian and D. Nowzari, "DNA Simulation of Nand Boolean Circuits," *AMO - Advanced Modeling and Optimization*, vol 6, No. 2, pp 33-41, 2004.
- [14] Y. Benenson, B. Gil, U. Ben-Dor, R. Adar R. and E. Shapiro, "An autonomous molecular computer for logical control of gene expression," *Nature*, vol. 429, no. 6990, pp. 423-429, 2004.
- [15] N. Dimitrova, "Intelligent Algorithms in Ambient and Biomedical Computing," *The Many Strands of DNA Computing*, vol. 7, pp. 21-35, 2006.
- [16] N. Nafiseh, Z. Sirous, L. Manigeh, K. Esmat, K. Morteza, "Rapi and sensitive detection of point mutations and DNA polymorphisms in Factor IX Gene by using the Single Strand Conformation Polymorphism (SSCP)," the fourth biotechnology congress, Krmansha, Iran, 2005.

- [17] S. Abdullah, "An improved DNA Computing approach using Heuristic Techniques," Ph.D. Thesis, Computer Sciences Department, University of Technology Baghdad, Iraq, 2008.
- [18] S. Hari S., K. Rajeev K. and S. Vikas, "An approach towards the solution of NP-Complete Problem," *Report and Opinion*, vol. 3, no. 5, 2011.
- [19] M. Adleman, "Computing with DNA, The manipulation of DNA to solve mathematical problems is redefining what is meant by computation," Copyright Scientific American, Inc, 1998.
- [20] M. Amos, "DNA Computing," Invited article for the *Encyclopedia of Complexity and System Science*, Springer, Manchester Metropolitan University, United Kingdom, 2008.
- [21] R. Sekhar, "DNA Computing-Graph Algorithms. Department of Mathematics," Indian Institute of Technology, This work is partially supported by Com2MaC-KOSEF, Korea; 2010.
- [22] G. Ibrahim, "Improving DNA Computing using Evolutionary Algorithms," Master Thesis. Software Engineering, College of Engineering, Salahaddin University, Hawler, Kurdistan, 2012.
- [23] M. Yamamoto, N. Matsuura, T. Shiba, Y. Kawazoe and A. Ohuchi, "DNA Computing, Solutions of Shortest Path Problems by Concentration Control," *Lecture Notes in Computer Science*, Springer link, 2002, vol. 2340, pp. 203-212.
- [24] J. Orlin, R. Ahuja, D. Simchi-Levi, S. Chopra, B. Golden and B. Kaku, "Networks and Flows," *Discrete Mathematics and Its Applications*, CRC Press; 1999.
- [25] T. Ootaa and Y. Yasuib, "Toric Sasaki-Einstein manifolds and Heun equations, *Nuclear Physics, Section B*, vol. 742, no. 1-3, 2006.
- [26] B. Datta and N. Nilakantan, "Two-dimensional weak pseudomanifolds on eight vertices," *Proceedings of the Indian Academy of Sciences - Mathematical Sciences*, vol. 112, no. 2, pp. 257-281, 2002.
- [27] A. Simovici and C. Djeraba, "Advanced Information and Knowledge Processing, Mathematical Tools for Data Mining, Set Theory, Partial Orders, Combinatorics," Springer-Verlag London Limited, 2008.

### APPENDIX A

AP	Represents the Average Path
DNA	Deoxyribonucleic Acid
drpPCR	Represents the number of dropped PCR solutions
drpSSCP	Represents the number of dropped SSCP solutions
E	Represents the number of network edges
EA	Evolutionary Algorithm
evolSSCPcross	Represents the number of Evolutionary SSCP Crossover operations
GA	Genetic Algorithms
GELP	Represents the number of Gel Electrophoresis solutions
GRS	Generate Random Solutions
HDNA	Heuristic Deoxyribonucleic Acid
HPP	Hamiltonian Path Problem
LIG	Represents the number of DNA Ligations
MC	Represents the Memory Capacity of the DNA Algorithm
MER	The length of the oligonucleotide is usually denoted by "mer" (from Greek meros, "part")
PCR	Polymerase Chain Reaction
pcrCross	Represents the number of PCR Crossover operations
pcrGen	Represents the number of PCR solutions generated by Crossover Operation
PCRP	Represents the number of PCR solutions
RNA	Ribonucleic Acid
RSG	Represents the number of Random Solutions Generated
SP	Represents the Shortest Path
SPP	Shortest Path Problem
SSCP	Single Strand Conformation Polymorphism, Represents the number of SSCP solutions
sscpCross	Represents the number of SSCP Crossover operations
sscpGen	Represents the number of SSCP solutions generated by Crossover Operations

# A Novel Paradigm for Symmetric Cryptosystem

Shadi R. Masadeh

Faculty of Information Technology  
Isra University  
Amman,  
Jordan

Hamza A. Al\_Sewadi

King Hussein Faculty of Computing  
Prince Sumaya University for  
Technology  
Amman, Jordan

Mohammad A. Wadi

Independent researcher  
Amman,  
Jordan

**Abstract**—Playfair cipher is the first known digraph polyalphabetic method. It relies on 5x5 uppercase alphabets matrix with simple substitution processes to be used for encryption and decryption. This paper proposes an enhanced variant of Playfair cipher algorithm that incorporates an algorithm for elaborate key generation starting with a seed accompanying the ciphertext and will be referred to as a Novel Paradigm for Symmetric Cryptosystem (NPSC).

The key generation, encryption and decryption processes implement modular calculations instead of the simple substitution used in the traditional Playfair cipher. It supports both alphabetic characters and numerals. This variant considerably enhances the security strength without increasing the matrix size as demonstrated by the experimentation results. Comparative studies of various critical factors with other reported versions of Playfair cipher and results were also included.

**Keywords**—*cryptography; security; symmetric systems; polyalphabetic cipher; key generation*

## I. INTRODUCTION

The first practical digraph substitution cryptosystem was Playfair cipher which is invented by Charles Wheatstone in 1854. It encrypts pairs of letters rather than single letters as in simple substitution cipher, hence reducing the hazard of frequency analysis attack significantly. Playfair cipher is reasonably fast and easy to be applied by low skilled persons, hence it was tactically used during World Wars I & II by all warring parties. Due to its powerful then and simple calculations, it proved suitable for the protection of non-critical information during combat. By the time the enemy cryptanalysts could break the message the information was useless to them [1].

The traditional Playfair cipher structure is a 5x5 matrix that contains the uppercase English alphabet with letters I and J are treated as equals. Uncounted books, papers and surveys included the structure, operation and characteristics of this ciphering technique were reported [2-8]. Due to its many peculiarities and shortcomings outlined by Srivastava and Gupta [9], many extended or modified versions of Playfair cipher schemes that would include alphanumeric and special characters were suggested in recent years as listed in section 2. However, this paper presents a new variant that incorporates modular arithmetic calculations rather than simple substitution and involves alphabetic and numerals text messages. It suggests a new key creation method out of a key seed for constructing the matrices for the encryption/decryption processes.

After the definition of Playfair Cipher in section 1, a brief survey of related work is given in section 2, followed by a detailed explanation of the proposed cryptosystem scheme in section 3. Section 4 includes the implementation of the algorithm and section 5 lists the experimental results and comparison with the original Playfair cipher and some of its variances. Finally section 6 concludes the paper.

## II. PLAYFAIR CIPHER VARIANTS

The most important variants of Playfair cipher will be chronically considered here.

Murali and Senthikumar [10] reported a rapid increase in security of the transmission over an unsecured channel by mapping random numbers to keyword of Playfair cipher. The corresponding numbers are then transmitted to the recipient instead of alphabetical letter.

Sastry et. al. [11] considered the 7 bits ASCII characters representation for the plaintext message characters denoted by codes from 0 to 127. Shannon's concept of confusion and diffusion was achieved by suitable variation in the traditional Playfair rules together with modification in the substitution tables.

Babu et. al. [12] implemented 6x6 matrix instead of 5x5 matrix in order to include number in their cipher, however, lowercase letters, white space and other printable characters cannot be handled. In order to include other uppercase letters and other characters, Srivastava and Gupta [9] modified their work by forming an 8 x 8 matrix. Beside they extended the technique by converting obtained cipher to the corresponding ASCII code values in decimal and further to corresponding 7 bits binary values, then applying Linear Feedback Shift Register to obtain the final ciphertext.

Agrawal et. al. [13] also used 5x5 matrix with consideration of letters I and J, but the frequency of digraph in the message is calculated first and the two letters with the lowest frequency were combined in the matrix immediately after the letters of the keyword.

Tunga and Mukherjee [14] extended the matrix to 16x16 in a multiple array structure in order to facilitate for information regarding spaces and the inclusion of "X" in the alphabet matrix. Besides it incorporated a shifting mechanism for rows and columns of the matrix to ensure that the encrypted text contains any ASCII code ranging between 0 – 255. Another 16x16 Matrix Playfair variant is also suggested by Dhenakaran and Ilayaraja [15].

Basu and Ray [16] implemented a rectangular 10x9 matrix in order to support almost all the printable characters including white space, and therefore increasing the size of the key domain to 90! and hardening the cryptanalysis tasks.

Hans et al. [17] implemented pattern swapping to produce a multiple key changes up to a maximum of fifty times. He adopted Random pattern of eight digits containing only numbers from 1-4.

Chand and Bhattacharyya [18] added to the 6x6 matrix of Babu et. al. [12] four iteration steps. They included letters and numbers from 0 to 9 without combining I and J in the same cell, hence user can write messages with all alphanumeric characters.

### III. THE NOVEL PARADIGM FOR SYMMETRIC CRYPTOSYSTEM (NPSC)

An enhanced cryptographic scheme (NPSC) is proposed in this work which is inspired by Playfair cipher that can encrypt alphanumeric messages. However it follows completely different and elaborate procedure. It adopts an elaborate key creation method and consists of two encryption/decryption algorithms and relies on modular arithmetic calculation for key generation and cryptographic processes. Two 5x5 matrices were employed as the backbone for this scheme, one for the alphabetic characters and the other for numerals. The key creation algorithm starts with a key seed that is chosen by the sender and attached to the encrypted message in order to produce these matrices, and then they will be used for message encryption at the sender side and decryption of the received cipher message at the receiver side. Details of all algorithms will be listed below.

#### A. Key generation Algorithm:

The key generation of the key starts with a seed that must consist of reasonable number both letters and numbers decided by the sender and embedded as a header for the ciphertext message. Its length is defined by certain agreement between the communicating parties. This seed is used independently by both sending and receiving parties to create the key according to properly designed process, and then produce two 5x5 matrices; one for the text and the other for the numerals. The first step for the key creation is to split the seed in two strings; text string and numerals string. Then text string is treated with algorithm-A of Fig 1 to create the first matrix and the numerals string is treated with algorithm-B of Fig 1 to create the second matrix.

**Algorithm-A:** for the text key string.

1. Open a file for the text string
2. Remove duplicate characters from the text file
3. Determine needed characters in the text file until length equals 4, 9 or 16
4. Invert text file
5. Remove duplicate letters from the text file.
6. Repeat step 3 until key length equals 4, 9 or 16
7. Place the text file in the lower right corner of 5x5 matrix using lower case letters
8. Subtract the text's characters from the alphabet, place the result into set  $X$

9. Place contents of  $X$  in the matrix using lower case letters in clock wise manner.

Number each matrix position sequentially and save it for later encryption and decryption

**Algorithm-B:** for the numeral key string.

1. Open file for numerical string
2. Convert each number into a letter using the alphabet array
3. Save the new text file
4. Remove duplicate letters from the text file
5. Add missing letters in the text file until length equals 4, 9 or 16
6. Invert text file
7. Remove duplicate letters
8. Add unused letters to the text file from the alphabet until length equals either 4, 9 or 16
9. Place the text file in the upper left corner of 5x5 matrix using upper case letters
10. Subtract the text's letters from the alphabet, place the result into  $X$
11. Place  $X$  in the matrix using upper case letters in clockwise manner

Number each matrix position sequentially and save it for later encryption and decryption

Fig. 1. The key creation algorithms

The created key length must be either equals 4, 9, or 16 characters. However, if the length of the key seed is not equal to any of these required number of characters, then some characters shall be added. For the algorithm-A, the added letters are determined by eq. 1, so, if the current key contains m characters, then next added character is determined by calculating its location value k first.

$$k = \left( \sum_{i=1}^m p_i + r \right) \bmod 26 \quad \dots \dots \dots (1)$$

where  $p_i$  is the location value of the  $i^{\text{th}}$  characters of the current key so far and  $r = m+1$ . The character matching the value of k is then added to the key string. This process is repeated until the required key length is achieved. For example if current key length is 7, then 2 more characters need are to be added. When the required key length of characters string is achieved, it is re-written in reverse order and checked for duplicate characters, if found, they are removed and replaced by unused letters from the alphabets in order. Moreover, in the case of obtaining the key with the required number of characters immediately from the message, there will be no need for any more calculations. It should be stated here that letters I and J are considered the same. Obviously, increasing the key length would enhance the security; hence it is recommended to add more characters to the key seed till the next required length is achieved.

A certain pointer is agreed upon by the communicating parties in order to distinguish the key seed from the ciphered

text. This can be either an integer number stating the length of the seed or an agreed upon separator between.

**B. Encryption Process**

To encrypt a message M which composes of both letters and numerals in the proposed NPSC cryptographic scheme, two possible algorithms are followed depending on the message contents whether it is text or numerical. The characters are encrypted sequentially and taken as ciphertext in the following procedure.

Each character is first identified; if it is a letter then algorithm 1 is used for encryption and if it is an integer number, algorithm 2 is used for encryption. The two algorithms are shown in Fig 2.

The resulting ciphertext string (C) content will be a mixture of lower case and uppercase letters corresponding to letters or numerals, respectively.

<p><b>Algorithm 1:</b> For letters encryption</p> <ol style="list-style-type: none"> <li>1. Read letter</li> <li>2. Read matrix A.</li> <li>3. Read letter's sequential position in alphabet, call it P</li> <li>4. Read letter's position in matrix A, call it X</li> <li>5. Calculate <math>T = (P+X) \bmod 26</math></li> <li>6. Read the letter having the value T in the matrix A and save it as lowercase ciphertext letter in C.</li> </ol> <p>Go to next character in the message M</p>
<p><b>Algorithm 2:</b> For numerals encryption</p> <ol style="list-style-type: none"> <li>1. Read numerical</li> <li>2. Read matrix file B</li> <li>3. Read number's sequential position, call it P</li> <li>4. Read number's position in matrix B, place it into X</li> <li>5. Calculate <math>T = (P+X) \bmod 26</math></li> <li>6. Read the letter which has the value T in matrix B, and save it as lowercase ciphertext letter in C.</li> </ol> <p>Go to next character in the message M</p>

Fig. 2. The encryption algorithms

To clarify how this process in executed, a detailed example is shown in the implementation section.

**C. Decryption Algorithm**

At the receiving end, the obtained message contains the key seed attached to the ciphered message. For the receiver to decrypt a ciphertext message first the key seed is identified and the removed from the received message in order to be used for creating the two matrices A and B. then the decryption process starts in order to recover the original plain text. Lowercase characters are treated by algorithm A and uppercase letters are treated by algorithm B. and since decryption is the reverse of the encryption process, hence subtraction is used instead of addition. It should be noted here that numbers might be negative; in such case, an addition of 25 is performed in order to acquire the original number.

**IV. IMPLEMENTATION**

To clarify the proposed NPSC variant, some examples will be included here for the use of the key seed to create the encryption/decryption matrices and to execute encryption and decryption processes. Two matrices are generated one for the letters and one for the numerals.

*Example 1:* matrices creation:

Suppose the key seed is the word “security5167”, then the seed for matrix A is “security” and that for matrix B is “5167”. The following will be followed for creating matrix A:

The alphabet is numbered first from 0 to 25, then seed letters are put in a table as in table I-a after removed duplicates, with the target to have 4 or 9 or 16 letters. In this table one extra character is required to get 9 letters. It is calculated in eq. 2.

TABLE I. DETERMINING THE KEY

s	e	c	U	R	I	t	y	_	r	y	t	I	u	c	e	s	_
1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

$18 + 4 + 2 + 20 + 17 + 8 + 19 + 24 + 9 \bmod 26 = 17 \dots (2)$

The value of 17 represents [r] in the alphabet, therefore the obtained key will be “security”

Now reversing the letters of this key and removing duplicates produces table I-b which again requires another letter.

Now repeating calculation similar to eq. 2, results into letter [v], hence the key becomes “rytiucsv”. It is used to filled the bottom left corner of a 5x5 matrix A, then this matrix is completed with remaining letters of the alphabet without duplication as described in section 3.1 using lowercase letters, see Fig. 3.

1	p	2	q	3	w	4	x	5	z
6	o	7	f	8	g	9	h	10	k
11	n	12	d	13	r	14	y	15	t
16	m	17	b	18	i/j	19	u	20	c
21	l	22	a	23	e	24	s	25	v

Fig. 3. Matrix A

The key seed for the numeral segment in this example is “5167”, which will be used for the creation of matrix B. After removing duplicate numbers the will look like table II-a, then replacing the corresponding letters from the alphabet produces table II-b.

TABLE II. DETERMINING THE KEY

5	1	6	7	F	B	G	H
1	2	3	4	1	2	3	4

Then after reversing the letters sequence of table II-b, the key will be “HGBF”. These letters will be placed in the upper right corner of a 5x5 matrix B, and will be completed with remaining letters of the alphabet without duplication as described in section 3.1 using uppercase letter, see Fig. 4.

1	H	2	G	3	A	4	K	5	R
6	B	7	F	8	C	9	L	10	S
11	I/J	12	E	13	D	14	M	15	T
16	Q	17	P	18	O	19	N	20	U
21	Z	22	Y	23	X	24	W	25	V

Fig. 4. Matrix A

Example 2: Encrypt the message M = “security5167”

Solution:

- s: 24 + 1 = 25 which is [v]
- e: 23 + 2 = 25 which is [v]
- c: 20 + 3 = 23 which is [e]
- u: 19 + 4 = 23 which is [e]
- r: 13 + 5 = 18 which is [i]
- i: 18 + 6 = 24 which is [s]
- t: 15 + 7 = 22 which is [a]
- y: 14 + 8 = 22 which is [a]
- 5 which equal F: 7 + 9 = 16 which is [Q]
- 1 which equal B: 6 + 10 = 16 which is [Q]
- 6 which equal G: 2 + 11 = 13 which is [D]
- 7 which equal H: 1 + 12 = 13 which is [D]

Therefore the ciphertext is “vveeisaaQQDD”.

### V. EXPERIMENTAL RESULTS

The proposed NPSC variant has been programmed in C++ language, and executed on different platforms with the following aspects in mind:

- Execution speed measurement for different package size under various operating environment (Windows 7, Windows Server 2012 and Linux).
- Transmission speed of encrypted messages over various network setups (wired and wireless networks).
- Comparison with playfair cipher.
- All results have been tested on the same environment using Intel Laptop core i7 processor with 8GB RAM

Therefore, the test included measurements of some important factors, such as CPU run time, power consumption, and the packet transmission time. All measurements were conducted for different package sizes; namely small package (1MB), medium package (10 MB) and large package (1 GB) for two different environments; namely wired network and wireless network. The results were listed in tables III –V and a comparison histogram is plotted for all studies cases in Figures 5-7.

Table III compares the measured factors of the proposed NPSC Cipher with the original Playfair cipher tested on a network using Intel Laptop core i7 processor with windows 7 operating system (OS), and then a histogram plot is drawn in Fig 5. It is shows that all measured factors have improved values in case of NPSC as compared with Playfair Cipher.

TABLE III. MEASUREMENT OF VARIOUS FACTORS FOR A NETWORK USING WINDOWS 7 OS

Algorithm Name	Packet Size	Wired Network			Wireless Network		
		CPU Time (ms)	Power Consumption (mw)	Transmission Time (ms)	CPU Time (ms)	Power Consumption (mw)	Transmission Time (ms)
Playfair	Small (1MB)	2.9	5.1	6.7	6.2	11.3	12.8
	Medium (10MB)	14.3	25.2	29.1	21.3	35.6	39.2
	Large (1GB)	20.4	35.4	40.3	31.6	42.1	50.6
NPSC	Small (1MB)	2.1	5.3	5.2	6.5	10.3	11.1
	Medium (10MB)	11.2	22.9	27	21	30	34.2
	Large (1GB)	18.7	33.2	39.5	30.2	41.6	49.1

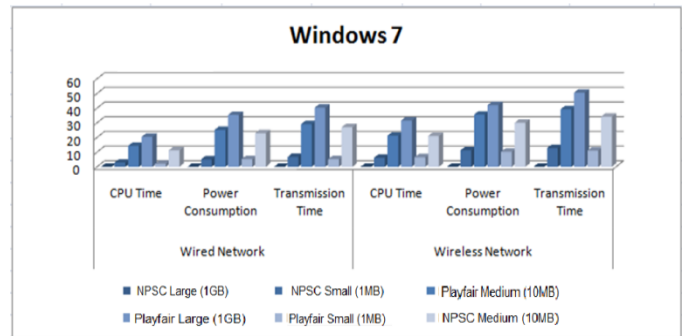


Fig. 5. Comparison of various factors for a network using windows 7 OS

Table IV compares the measured factors of the proposed NPSC Cipher with the original Playfair cipher tested on a network using windows Server 2012 operating system, and a histogram plot is drawn in fig 6. It is also shows that all measured factors have improved values in case of NPSC as compared with Playfair Cipher.

TABLE IV. MEASUREMENT OF VARIOUS FACTORS FOR A NETWORK USING WINDOWS SERVER 2012 OS

Algorithm Name	Packet Size	Wired Network			Wireless Network		
		CPU Time (ms)	Power Consumption (mw)	Transmission Time (ms)	CPU Time (ms)	Power Consumption (mw)	Transmission Time (Mbps)
Playfair	Small 1MB	3.7	10.6	9.5	7.4	17.2	12.5
	Medium 10MB	16.9	26.1	35.9	24.2	38.4	40.2
	Large 1GB	23	45.6	50.6	37.6	52.1	59.2
NPSC	Small 1MB	3.5	10.0	7.4	6.5	16.1	12
	Medium 10MB	15.5	27.6	30.2	20.3	34.6	34.7
	Large 1GB	22.7	43.3	45.3	33.1	50.2	51.1



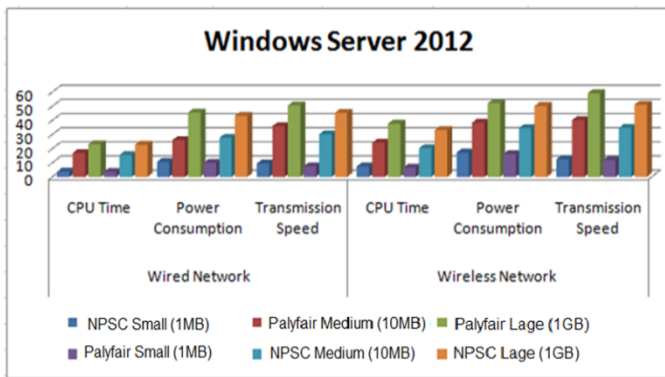


Fig. 6. Comparison of various factors for a network using windows Server 2012 OS

Table V compares the measured factors of the proposed NPSC Cipher with the original Playfair cipher tested on a network using Linux operating system, and a histogram plot is drawn in fig 7. Again it is shows that all measured factors have improved values in case of NPSC as compared with Playfair Cipher.

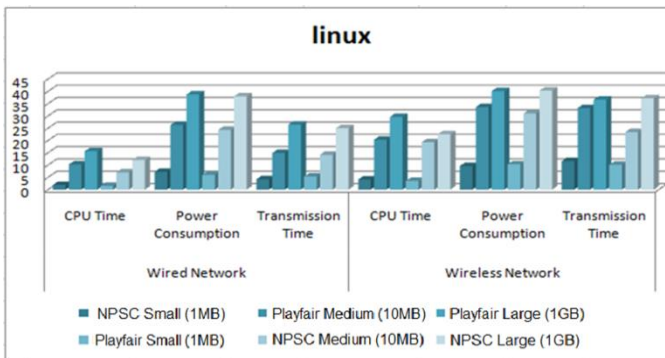


Fig. 7. Comparison of various factors for a network using Linux OS

TABLE V. MEASUREMENT OF VARIOUS FACTORS FOR A NETWORK USING LINUX OS

Algorithm Name	Packet Size	Wired Network			Wireless Network		
		CPU Time (ms)	Power Consumption (mw)	Transmission Time (Mbps)	CPU Time (ms)	Power Consumption (mw)	Transmission Time (Mbps)
Playfair	Small (1MB)	1.9	7.2	4.2	4.1	9.6	11.6
	Medium (10MB)	10.2	26.3	14.9	20.3	33.6	33.2
	Large (1GB)	15.6	38.7	26.4	29.6	40.1	36.6
NPSC	Small (1MB)	1.5	6	5.3	3.5	10.3	10.1
	Medium (10MB)	7	24.3	14.2	19.3	31	23.4
	Large (1GB)	12	38	25	22.5	40.3	37.2

Despite the elaborate computation involved in the proposed NPSC design which includes modular mathematics rather than only substitution as for the original Playfair cipher, the observed empirical results have all demonstrated improvements in algorithm execution time, power

consumption and transmission speed over computer networks for different platforms.

The security of NPSC can be also compared with variant of playfair cipher, as shown in table VI.

### VI. SECURITY ISSUES OF NPSC CIPHER

Any cryptosystem is designed to stand cryptanalysis attacks relying on time complexity and space complexity. The original Playfair Cipher is vulnerable to Brute force attack but has reasonable resistance to frequency analysis attack. For the NPSC cipher, these two attacks may be considered here.

#### A. Brute Force Attack [2]:

This attack systematically attempts all possible key combination; hence, larger the key space results into more secure cipher. In NPSC cipher, two independent 5x5 matrices were for encryption and decryption purpose. Therefore, the key space for building these matrices will be (26x26)x(26x26) resulting into 456976 different possible combinations for the Brute force attack as compared with 26x26 diagrams for Playfair cipher.

#### B. Frequency Analysis

Frequency analysis is the study of the frequency of occurrence for letters or groups of letters in a ciphertext [20]. It is based on the statistical fact that, each letter or group of letters for any language has certain frequency of occurrence. This frequency would be decided with more accuracy if the written text was of considerable size. The characteristic distribution of these letters is found to be almost the same for any stretch of text of reasonable length [21]. However, the proposed NPSC cipher relies mathematical calculations and not digraph substitution; a thing put the frequency attack out of the question.

#### C. Comparison Analysis:

A comparison study is conducted for the proposed NPSC cipher in contrast with the original Playfair cipher and some of its reported variants and listed in table VI [8]. It includes the key space that is available for Brute force attack, the number of possible diagrams need to be searched for ciphertext only attack, and the probability of occurrence of an element for frequency analysis attack.

TABLE VI. PLAYFAIR VARIANT COMPARISON

Playfair Cipher	Key space for Brute force attack	Number of diagrams to be searched for ciphertext only attack	Probability of occurrence of an element for frequency analysis attack
Original	25!	676	0.038
Srivastava & Gupta [9]	54!	4096	0.016
Babu et al [12]	86!	1296	0.028
Hans et al [17]	26!*24*24	Difficult	Difficult
Chad et al [17]	86!	1296	0.028
Verma et al [19]	54!	4096	0.016
Proposed NPSC	86!	456976	Difficult

## VII. CONCLUSIONS

This paper introduced few variations to Playfair ciphers such as using numerals together with letters, creating the encryption/decryption key from an alphanumeric seed, and the expansion of the encryption/decryption matrices to two instead of one matrix. Also the encryption/decryption processes are performed using modular arithmetic.

These added values to the ciphering technique have given larger key domain size for brute force attack and increased number of diagrams needs to be searched for ciphertext only attack and also handicapped the frequency of occurrence analysis attack. More work is needed to be pursued to study the avalanche effect on this cipher.

### REFERENCES

- [1] Wikipedia ([http://en.wikipedia.org/wiki/Playfair\\_cipher](http://en.wikipedia.org/wiki/Playfair_cipher)), last visited Oct. 2015.
- [2] W. Stallings, "Cryptography and network security: Principle and practice", 5<sup>th</sup> ed, Pearson Education, 2011.
- [3] D. Bruff, "The Playfair cipher revealed wynn", MLAS 280-07 Cryptography July 13, 2009.
- [4] E. Baldwin and D. Bruff, "Playfair cipher", FYWS Cryptology October 27, 2010. <<http://derekbruff.org/blogs/fywscrypto/files/2010/11/Baldwin-Essay-2.pdf>>
- [5] K. R. Babu, S.Udaya Kumar and A.V. Babu, "A survey on cryptography and steganography methods for information security", International Journal of Computer Applications (0975 – 8887), Vol. 12, No.2, November 2010,
- [6] "The Playfair algorithm description", <http://macliang.acns.carleton.edu/falk/other/playfair.htm>.
- [7] M. Kumar, R. Mishra, R. K. Pandey and P. Singh, "Comparing classical encryption with modern techniques", proceedings of S-JPSET, Vol. 1, No. 1, 2010.
- [8] P. Goyal, G. Sharma and S. S. Kushwah, "Network security: A survey paper on Playfair cipher and its variants", International Journal of Urban Design for Ubiquitous Computing Vol. 3, No.1, 2015, pp.1-6. <http://dx.doi.org/10.14257/ijuduc.2015.3.1.01>
- [9] S. S. Srivastava and Nitin Gupta, "Novel approach to security using extended Playfair cipher", International Journal of Computer Applications (0975 – 8887), Vol 20, No.6, April 2011.
- [10] P. Murali and G. Senthilkumar, "A modified version of Playfair cipher using linear feedback shift register", International Journal of Computer Science and Network Security (IJCSNS), Vol.8, No.12, December 2008.
- [11] U. Sastry, N. R. Shankar and S. D. Bhavani, "A modified Playfair cipher involving interweaving and iteration", International Journal of Computer Theory and Engineering, Vol. 1, No. 5, December, 2009, 1793-8201.
- [12] K. R. Babu, U. Kumar, A. V. Babu, I.V.N.S Aditya and P. Komuraiah, "An extension to traditional Playfair cryptographic method", International Journal of Computer Applications, (0975 – 8887), Vol.17, No.5, March 2011.
- [13] G. Agrawal, S. Singh and M. Agarwal, "An enhanced and secure Playfair cipher by introducing the frequency of letters in any plain text", Journal of Current Computer Science and Technology Vol. 1, No. 3, 2011, PP10-16.
- [14] H. Tunga and S. Mukherjee, "A new modified Playfair algorithm based on frequency analysis", International Journal of Emerging Technology and Advanced Engineering, Vol. 2, No. 1, January, 2012.
- [15] S. S. Dhenakaran and M. Ilayaraja, "Extension of Playfair cipher using 16X16 matrix", International Journal of Computer Applications (0975 – 8887) vol. 48, No. 7, June, 2012.
- [16] S. Basu and U. K. Ray, "Modified Playfair cipher using rectangular matrix", International Journal of Computer Applications (0975 – 8887) Vol. 46, No.9, May, 2012.
- [17] H. Hans, R. Johari and V. Gautam, "An extended Playfair cipher using rotation and random swap patterns," 5th IEEE International Conference on Computer and Communication Technology, 2014.
- [18] N. Chan and S. Bhattacharyya, "A Novel approach for encryption of text messages using Playfair cipher 6 by 6 matrix with four iteration steps", International Journal of Engineering Science and Innovative Technology (IJESIT), Vol. 3, No. 1, January, 2014.
- [19] V. Verma, D. Kaur, R. K. Singh and A. Kaur, "3D- Playfair cipher with additional bitwise operation", Control Computing Communication & Materials (ICCCCM), 2013 International Conference on IEEE, (2013), August, 2013, PP1-6.
- [20] R. P. Dhiren, "Information security theory and practice", 1<sup>st</sup> ed, Prentice-Hall of India Private Limited, 2008.
- [21] Harrison K., B. Munro and T. Spiller, "Security through uncertainty", P Laboratories, February, 2007.

# Integrating Semantic Features for Enhancing Arabic Named Entity Recognition

Hamzah A. Alsayadi

Ibb University, Yemen  
MSc student in Faculty of Computers and  
Information, Cairo University  
Cairo, Egypt

Abeer M. ElKorany

Dept. of Computer Science, Faculty of Computers and  
Information, Cairo University  
Cairo, Egypt

**Abstract**—Named Entity Recognition (NER) is currently an essential research area that supports many tasks in NLP. Its goal is to find a solution to boost accurately the named entities identification. This paper presents an integrated semantic-based Machine learning (ML) model for Arabic Named Entity Recognition (ANER) problem. The basic idea of that model is to combine several linguistic features and to utilize syntactic dependencies to infer semantic relations between named entities. The proposed model focused on recognizing three types of named entities: person, organization and location. Accordingly, it combines internal features that represented linguistic features as well as external features that represent the semantic of relations between the three named entities to enhance the accuracy of recognizing them using external knowledge source such as Arabic WordNet ontology (ANW). We introduced both features to CRF classifier, which are effective for ANER. Experimental results show that this approach can achieve an overall F-measure around 87.86% and 84.72% for ANERCorp and ALTEC datasets respectively.

**Keywords**—Arabic Named Entity Recognition (ANER); Conditional Random Fields (CRF); Domain Ontology; Semantic Relation Feature (SRF); Arabic WordNet ontology (ANW)

## I. INTRODUCTION

Named Entity Recognition (NER) was introduced in 1990 at the Message Understanding Conferences (MUC-6) [1]. NER is one task of an information extraction to classify proper names from raw texts into types of names [1]. Three major tasks of NER were covered: (person, location, and organization) called ENAMEX, (temporal expressions) called TIMEX, and (some numerical expressions such as monetary amounts and other types of units) called NUMEX [1, 5]. There are other Named Entities (NEs) were defined by NER such as biology domain (like gene, DNA, and RNA NEs), Behavioral Health like (healthy food), and biomedical like (diseases NE) [10 - 12]. In this paper, we deal only with ENAMEX. The goal of Named Entity Recognition (NER) task is the enhancing the accuracy concerning the named entities recognition and extraction [3]. NER task is important for many natural language processing applications such as Search results clustering, Machine Translation, Navigation Systems, enhancing Information Retrieval, and Improving results in Question Answering [1,4].

Due to the political and economic importance of the Arabic language, in the last decades, the NLP researchers started to get interest in research fields in the Arabic language such as Arabic Named Entity Recognition (ANER) [5]. The Arabic language has a rich vocabulary, morphology, and syntax; also, it has a complex morphology [1, 6]. The Arabic language has three styles, Classical Arabic (CA), Modern Standard Arabic (MAS) and Colloquial/Dialects Arabic (DA) [5]. In this work, MSA are dealt. There are challenges for ANER such as high morphological ambiguity, complexity and common noun/words ambiguities. The researchers in ANER tried to collect appropriate data to include all possible language cases having these characteristics and peculiarities such as ANERCorp<sup>2</sup> and ALTEC<sup>3</sup> datasets. Also, they developed tools for that data such as MADA<sup>4</sup>, Stanford POS Tagger<sup>5</sup>, and AMIRA<sup>6</sup>.

The ANER researcher developed system depends on two approaches Ruled Based approach [7, 8, 9, 13, 14] or Machine learning (ML) approach [15 - 18]. The systems were built using Ruled-Based approach, which depends on linguistic rules for recognizing NEs. These rules are usually regular expressions or finite-state transducers. The advantage of the rule-based NER systems is that they are depend on the core of linguistic knowledge. However, any update or maintenance required for these systems is time-consuming and labor-intensive; also, it requires full knowledge of the language [1]. ML approach is to learn NE tagging decisions from annotated texts. The most common approach that is used in ML for NER is Supervised Learning (SL). It represents the NER problem as a classification task that distinguishes between different types of names entities. The advantage of ML-based NER systems are the ease of maintenance, modifications, and adaptation over time. According to [4, 15], CRF and SVM had been proven as the best techniques for ANER. The researchers in [4] proved that CRF is better than other techniques while in [15] they did not state whether CRF is better than SVM or not in Arabic NER.

In this paper, an integrated semantic-based ML is applied for ANER. CRF is used CRF as the classification engine for recognizing three named entity (NE) classes; person, location,

<sup>1</sup> <http://cs.nyu.edu/cs/faculty/grishman/muc6.html>

<sup>2</sup> <http://users.dsic.upv.es/~ybenajiba/>

<sup>3</sup> [http://www.altec-center.org/Repository\\_65.html](http://www.altec-center.org/Repository_65.html)

<sup>4</sup> <http://www1.ccls.columbia.edu/~cadim/MADA>

<sup>5</sup> <http://www-nlp.stanford.edu/software/tagger.shtml>

<sup>6</sup> <http://www.cs.columbia.edu/~mdiaib/>

and organization names. This integration is new for ANER, to the best of our knowledge, since it has not been utilized in ANER. This model combines internal features that represented linguistic features as well as external features to represent the semantic of Arabic language. Arabic semantics relate primarily to the semantic correlates of morphological patterns. This correlation is extracted from two different resources each represented the relationships that could exist between the extracted named entities such as ontology (Arabic wordNet ontology) in form of classes, instances, and relations between entity classes, and feed it to CRF classifier as a set of features to enhance the classification process. These semantic features are efficient for ANER, and over performed other CRF that used less number of features with better accuracy.

The paper is organized as follows: Section 2 illustrates some of Arabic Language challenges; Section 3 gives an overview of the domain ontology; Section 4 explain some of the previous systems as related work; Section 5 discusses the components of architecture system; Section 6 show the data that used in this system; Section 7 contains an Evaluation Criteria; Section 8 includes an experimental; Section 9 submits results and discussion Finally; Section 10 present the conclusion and future work.

## II. ARABIC LANGUAGE CHALLENGES

We focus on Arabic NER that has several challenges and characteristics:

1) *Lack of capital letters*: A named entity in Latin languages is usually distinguished by a capital letter at the word beginning. However, Arabic lack the capital letter, so the detection of NE in text based on the letters case more difficult. The lexical triggers used to overcome this problem, which has used that are derived from analyzing the surrounding context of NEs while some others researchers have used the English translation of the NE what is known as the glossing feature produced by the MADA tool [2, 3, 22].

2) *Complex Morphology*: the Arabic language has a complex morphology due to the agglutinative nature of language. Agglutinative morphemes have three types: stems, affixes, and clitics. The stem is the primitive form of the word. Affix letters are usually added to the stem, which has three types: prefixes attached to beginning of the stem, suffixes attached to end of the stem, and circumfixes that surround the stem. Clitics are also added to the stem after affixes. Clitics are either proclitic that come before the word or enclitics that come after the word. The conjunction “و” (waw, and) and object pronoun “هن” (hm) are examples of proclitic and enclitics, respectively. A more general example is the word “وسيدرسونها” (and-they-will-study-it) [1].

3) *Ambiguity*: Arabic text has the different meaning for one word (Ambiguity). For example (رغاب/Ragab) in Arabic may be used as a person name, and month. The word diacritization is important factor for word meaning, for example, (قطر) which if it is diacritized as قَطْر it means country Qatar but if it is diacritized قَطْر it means Diameter or territory [1, 2, 22].

4) *Arabic is a high inflectional language*; often a single word has more than one affix such that it may be expressed as a combination of prefix(s), lemma, and suffix(s) as Word = prefix (es) + lemma + suffix (es). The prefixes are prepositions, conjunctions, or articles. The suffixes are generally personal/possessive or objects anaphora. For example, the Arabic word “ويعروبتنا” is interpreted in English as “and with our Arabism” [2, 3].

5) *Writing Styles Arabic (Spelling variants) has a high level of typographic forms and ambiguity spelling*: An NE can be writing in a many of ways. This multiplicity arises from both different ways of writing the Arabic writers and ambiguous form of transcription schemes. There is no fixed standardization for writing the word like English. For example, the word ‘جرام’, jrAm1, ‘Gram’, can also be written as ‘غرام’, grAm, with the same meaning, also the word جوجل / Google can be written as غوجل, other example the word سوريا / Syria can be written as سورية [1, 2, 22].

6) *Systematic Spelling Mistakes* Typographical errors were frequently made by Arabic writers according to certain characters. For example, الإسلامي / The Islamic with (ل) can be written الإسلامي with (ل), and العربية / the Arabia with (ة) can be written as العربية with (ة) [1].

7) Some foreign persons’ names when it was transliterated into Arabic could be identified as pronouns or prepositions such as [Ho, Anna, Ann, and, Lee] their different pronouns or prepositions are [He, I, That, Mine] [22].

8) *Lack of Resources*: Large collections of tagged documents (corpora), gazetteers (predefined lists of typed NEs), and NLP tools, are either rare or not free. This challenge makes collecting and analyzing the data is time-consuming particularly if the NER technique depends on such resources [23, 24]. There are few corpora such as the free ANERCorp, the commercial ACE (2003 – 2005)<sup>7</sup> and ALTEC.

## III. DOMAIN ONTOLOGY

Domain ontology, as a formal specification of a shared conceptualization, defines the Knowledge base of the concept, attributes, relations between concepts and properties even relations between properties. Moreover, it describes axioms, individuals and relations between them, and provides sharing knowledge. It has a better capacity of semantic Interpretation. The specific domain wordNet ontology was used.

WordNet is a large lexical database for English by Princeton. It contains information about 147,278 words divided into nouns, adjectives, verbs, and adverbs. Then the words are expand divided into 206,941 senses, with an average of 1.4 senses per word. These senses are grouped by synonymy into 117,659 unorganized sets called synsets. Words in the same synset refer to the same concept and are may be used in many contexts mutually. There are also some semantic relations between the synsets such as the hyponym and hypernym relations. Thus, WordNet is sometimes considered as a lexical ontology. WordNet has realized great

<sup>7</sup> <https://catalog.ldc.upenn.edu/LDC2006T06>

success and became the dominant English lexicon in NLP applications [25].

The Arabic WordNet (AWN) is a wordNet for Modern Standard Arabic (MSA), it was built depend on the design and contents of WordNet (WN) [26]. Thus, the AWN synsets are linked to WN synsets directly. Up to now, AWN consists of 13,808 non-diacratized Arabic words divided into 23,481 senses that form 11,269 Arabic synsets. All of the synsets are connected to the corresponding English synsets in WordNet. The low AWN/WN ratios suggest low coverage of Arabic words in AWN, which can be easily verified as some of the commonly used Arabic words are missing, such as the noun 'بطولة' (championship), the verb 'تقابل' (meet), and the adjective 'أفريقي' (African) [25].

#### IV. RELATED WORK

There are many researchers that applied ML-based for ANER in order to learn NE tagging decisions from annotated texts. There are techniques utilized for ANER are Support Vector Machines (SVM), Conditional Random Fields (CRF), Maximum Entropy (ME), artificial neural network (ANN), and Decision Trees. Each technique need features for NEs identification such as gazetteers features, POS tags and morphology features. Some researchers used CRF [4, 16, 19] depend on their features while in [20] used SVM, other researchers used ME [21]. Finally, in [3], the authors used ANN. Other researchers used CRF and SVM in [15, 17].

In [16] authors introduced a system for improving NER on microblogs, it contain three methods: (1) using large gazetteers from Wikipedia, (2) domain adaptation, and (3) a two-pass semi-supervised method. They used CRF classifier (CRF++<sup>8</sup>). They tagged new training set from Tweeter. The evaluation of system depended on ANERcorp and new training set (which they tagged). They compared their system with other systems; this system shows an improvement of 35.3 F-measure points over other systems.

The authors [17] proposed a simplified feature set system. This system dealt with only some of the Arabic morphological and Arabic orthographic complexities features. They used CRF classifier to identification NEs. They evaluated their work using ANERcorp and ACE2005 dataset. The result of the system proved the effectiveness of simplified feature set for ANER.

In [18] the authors developed system using Cross-lingual Features. They used three Arabic and English Wikipedia cross-language links Cross-lingual Capitalization, Transliteration Mining and using DBpedia. The work used CRF, was evaluated using ANERcorp dataset for training and testing, also used NEWS Test Set and TWEETS Test Set. In this work, the authors showed how cross-lingual Features enhanced ANER.

Semi-supervised learning was used in [27] to develop ASemiNER, a semisupervised algorithm for identifying Named Entities (NEs). The system including Pattern Induction, Instance Extraction, and Instance Ranking/Selection Methodology. ASemiNER does not require

any annotated corpora or any gazetteers, but it was compared with ANERcorp and ACE2005 dataset.

In [15] the authors investigated a large of features sets in order to get the optimal feature sets. Multiple classifiers were used in this system to recognize NEs SVM and CRF. They ACE 2003, ACE 2004 and ACE 2005 data sets. The multi-classifier and language independent features outperform the system in [20] that used one classifier by 0.79 F-measure.

In [3] the authors developed the system using Artificial Neural Networks (ANN) approach. The system including three processes preprocessing of the data, transforming the Arabic letters to Text Romanization and applying the ANN classifier to the text. They used ANERcorp dataset and data collected manually from diverse web sources for evaluation. The authors compared the result of the system between decision trees and ANN approaches. The result demonstrated the ANN achieves higher results than that to get from the decision trees approach.

In [2], the authors presented a solution for ANER; this solution is an integration between two machine learning approaches, bootstrapping semi-supervised pattern recognition and CRF classifier as a supervised technique. This system including three modules CRF classifier, pattern recognizer, and the matcher module. In this solution is used RDI-ArabSemanticDB tool and RDIArabMorpho-POS tagger. They used ANERcorp (person, location, and organization) and crawled from the web other NEs for the system evaluation. This integration is designed to increase the CRF F-measure.

In [4], the authors developed their previous works (ANERSys) in [21] to enhance the accuracy of this system using Conditional Random Fields. The performance results achieved on ANERcorp dataset. They identify Person, Location, and Organization classes with F-measure of 73.34%, 89.74%, 65.76%, and 61.47% respectively. They prove that CRF achieves the result better than their previous work in [21] by 12 points in the F-measure average of all classes.

In [22] the authors developed their system in [28]. They used integrated approach: a) name dictionaries and b) name clusters with a statistical model based on extracting patterns that indicate the existence of person's names. They used list of names more than list in [28] was named full\_names\_19000\_list. The result in this system is better than their previous work [20] by 4.09 F-measure.

In [29] the authors used semi-supervised and distance learning techniques, then Bayesian Classifier Combination (BCC) to recognize Arabic NEs. They built Wikipedia-derived corpus (WDC). They used the dataset that built and ANERcorp dataset for evaluation. Previous Systems perform better than this work.

For the best of our knowledge, there is no work used semantic information in ontology such as semantic relations in Arabic Named entity recognition (ANER). The semantic information in ontology was utilized in this work, but there are systems in other languages implemented it with some differences.

<sup>8</sup> <https://taku910.github.io/crfpp/#download>

In [10] the authors used semantic information in ontology. They used internal features (POS and Word n-gram) and external features from ontology using notebook domain ontology. The CRF classifier was utilized in this system. The system evaluated using ChnSentiCorp corpus.

The ontological features in [31] used for Vietnamese named entity recognition (NER). The authors used CRF classifier and VN-KIM dataset.

In [11] the authors proposed system for Recognize Named Entity in Behavioral Health. They built the manual ontology. The specific domain was used in this system using wordNet ontology.

The three last works showed the advantage of adopting semantic information in the ontology for NER.

### V. THE SYSTEM ARCHITECTURE

In this work, ML-based ANER is proposed which utilizes two types of features: a) internal features, b) external features. Figure (1) illustrates the Architecture of this system. This system includes four phases: preprocessing (prepare and clean data), features extraction, training, and testing.

#### A. Pre-Processing

This step includes cleaning the data such as splitting the sentence and tokenization. For preparing and cleaning data, the following processes was applied [31]:

- Remove the Redundant space among the words; remove all characters and symbols that attached to the word from the corpus such as (-, \*, +, etc.) [4].
- Omit the prefixes and suffixes that attached as the conjunction ((wa / و) and (ba/ ب)).
- Remove the preposition li (lam) (ل).
- Delete all diacritics within the text.

Splitting the sentence is the task of segmenting the text into the sentences. The goal of this step is to define the boundaries of phases in the text according to POS tagger. The tokenization is the process that analyses and splits the input text into tokens such as, word, number, symbol, space. The objective of this step is to divide the sentence into the tokens in order help us to extract the features from ontology. In this step, the white space characters was used to define the tokens in the sentence.

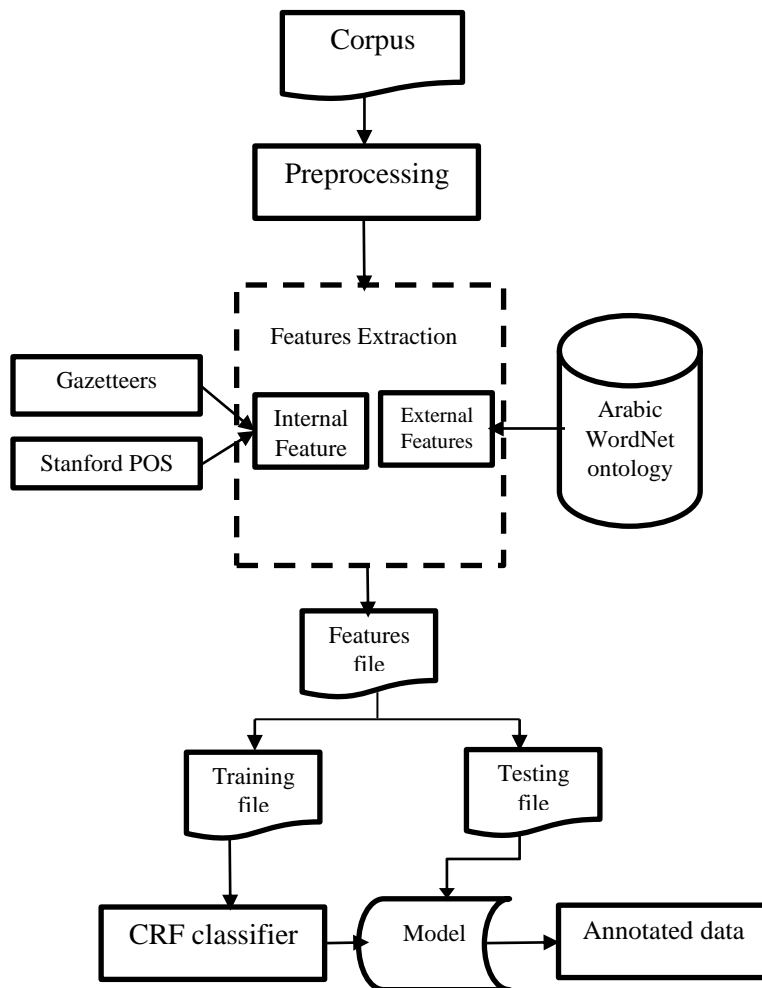


Fig. 1. The system Architecture

### B. CRF classifier

In this work, a modified Conditional Random Fields (CRF) is applied. CRF as described in [32] is a probabilistic framework used for segmenting and labeling the sequential data. It is a generalization of Hidden Markov Model in which its undirected graph contains nodes to represent the label sequence  $y$  corresponding to the sequence  $x$ . CRF finds the label which maximizes the conditional probability  $p(y|x)$  for a sequence  $x$ . The following equations represent CRF model:

$$p(c|y) = \frac{1}{Z(x)} * \exp\left(\sum_i \lambda_i f_i(x, c)\right) \quad (1)$$

$$Z(x) = \sum_{c'} \exp\left(\sum_i \lambda_i f_i(x, c')\right) \quad (2)$$

Where  $c$  is the class,  $x$  is a context information and  $f_i(x, c)$  is the  $i$ th feature.

In any ML approach for NER, there are two steps, training, and testing, as shown in figure (1). The first one builds the classifier model by using a set of features. In the second step, the classifier model that was built by the training step is utilized to predict a class for each token (word).

### C. Features sets

In the proposed framework, there are two category of features into two types: a) internal features, and b) external features that

a) **Internal features** some important features for Arabic text are introduced as following:

- 1) **Word (WF)**: is the word itself.
- 2) **Part of speech features (POS)**: part of speech tag is useful for ANER for determine the noun. The Stanford POS Tagger<sup>9</sup> was utilized to extract many tags NNP, NN, IN, JJ, NNPDT, NNDT, ... etc.
- 3) **Gazetteers features (GAZ)**: external resources and classes in Corpus are used, which are mentioned in data collection in section 5 to represent the existence of the word in the gazetteers.
- 4) **Indicator features (CF)**  
Indicator features are one of the most important features that lead to enhance the accuracy of NER recognition as it support the usage of semantic field feature [2]. They represents a set of words that may be used to identify NE such as preceding indicator words and post indicator. These words are used to recognize some names. For example, (الرئيس | الرئيس) (the President), (السيدة | السيدة) (Mrs. | Miss), and (أبو) (Abu) for person names, (دولة | Country), (مدينة | City), and (شارع | Street) for location names, and (مجموعة | Group), (هيئة | Organization), (نادي | Culp), (شركة | Company), and (بنك | Bank) for Organization names.

#### 5) Gram character features (GF)

These features Presents the first/last two and three letters of the word. This feature is very important for ANER. For

example, (عبد | Abd) is very repetitive prefix in Arabic person names.

#### b) External features (ontology or semantic features)

AWN tool<sup>10</sup> has been modified to be able to analyze texts of wordNet ontology and establish correspondence between syntactic dependencies and semantic relations in order to extract the following features:

1) **Class feature**: represents the ontology's concept for the token as person for person names, (city | country or location) for location names, and (company or organization) classes for organization names.

2) **Instance feature**: which represent the corresponding instance for NE's token.

3) **Relation features** these features represent the relations between each two named entities. Therefore, in this step we aim to identify the trigger words that express the semantic relations between NEs from Arabic text. Based on the probability of relation that could exist between pairs of named entities (person, organization, location), cross multiplication is applied and we extracted all possible combination that may appear in the ontology. Furthermore, since Arabic relations could appear before the first NE, between NEs or after the second NE [33] such as (أبو تريكة لعب للأهلي) or (لعب أبو تريكة للأهلي) [Abu Trika **played** for Al Ahli]. In this work we only focus on the relation between a pairs of NEs such as a relation between person's concept and location's concept for example (بارك أوباما) [Barack Obama, **the President of** the United States] (Obama) is person name, (United States) is location name and (the President of) the relation between them. The relation between the person's concept and organization's concept for example (بيل جيتس مالك شركة مايكروسوفت) [Bill Gates, **owner of** Microsoft Company.]. The relation between location's concept and organization's concept for example (نادي برشلونة من اسبانيا) [FC Barcelona **from** Spain]. Finally relation between the location's concept and themselves for example (القاهرة عاصمة مصر) [Cairo is **the capital of** Egypt]. Accordingly, two types of relations are identified:

a) **Explicit relations** which explicitly identified by Arabic wordNet (AWN) and are targeting the following pairs (PERS-PERS, LOC-LOC, PERS-ORG, ORG-LOC, and PERS-LOC)

b) **Semantic Relations** which are extracted depending on relationship between classes and their properties in AWN. Those types of relations are used to identify the following pairs (PERS-ORG, PERS-LOC, ORG-LOC, and LOC-LOC). The following algorithm shown in figure (2) is developed, to identify those possible relations between pairs of names entities. The algorithm works as follows:

- 1) For each sentence in the corpus, each two tokens are recognized and their classes are identified.
- 2) If both tokens are not belonging to the same class, calculate the semantic distance (SD) [11], which is

<sup>9</sup> <http://www-nlp.stanford.edu/software/tagger.shtml>

<sup>10</sup> <http://sourceforge.net/projects/awnbrowser/>



considered as the distance in hypernym/hyponym tree between the two classes of tokens.

- 3) If the semantic distance (SD) less than 3 and greater than 1 there is a relation between two tokens.
- 4) else if both classes are not found in AWN or both tokens have the same class, then there is no relation

```
Initialize R= O // Represent the relation between classes
Take two tokens from sentence
Find the classes for two tokens from AWN
CT1, CT2 // Represent the classes of token1, and
token2 respectively
IF CT1 and CT2 not found OR CT1= CT2 Then
Return O
Else
SD=Calculate the semantic distance between CT1 and CT2
IF 1 ≤ SD ≤ 3 Then
R= REL
Return R
Else
Return O
End IF
End IF
```

Fig. 2. Semantic relation extraction algorithm (SREA)

## VI. DATA COLLECTION

In order to train, and test the proposed ANER, necessary linguistic resources of different main categories were used: corpus, gazetteers, dictionaries, and AWN. Two corpuses are used for training, and testing the system. In this section, a description of all linguistic resources is presented.

1) *ANERcorp*<sup>11</sup> dataset, which is freely available for research purposes, is a corpus prepared by Yassine Benajiba in ANER. It has 4901 sentences with 150286 tokens. Each token in this corpus is tagged according to the following classes:

- B-PERS: The Beginning of the person name.
- I-PERS: The Inside of the person name.
- B-LOC: The Beginning of the location name.
- I-LOC: The Inside of the location name.
- B-ORG: The Beginning of the organization name.
- I-ORG: The Inside of the organization name.
- O: The word is not a named entity (Other).

2) *ALTEC*<sup>12</sup> dataset which is not free, is a corpus prepared by Arabic Language Technology Center, it has 288737 tokens. Each token in this corpus is tagged according to the following classes:

- B-nep: The Beginning of the person name.
- I-nep: The Inside of the person name.
- B-nel: The Beginning of the location name.
- I-nel: The Inside of the location name.

- B-neo: The Beginning of the organization name.
- I-neo: The Inside of the organization name.
- O: The word is not a named entity (Other).

### 3) Gazetteers

Different gazetteers are integrated such as:

**ANERGazet**<sup>13</sup> is prepared by Yassine Benajiba gazetteers contained 2305 person names, 1785 location names and 390 organization names.

**Gate gazetteers** were containing 1883 person names, 403 location names and 215 organization names

**Lists of names**<sup>14</sup> form Wikipedia gazetteers were containing 16037 person names, and 4857 location names.

### 4) Arabic Wordnet<sup>15</sup>

The AWN ontology contains a large amount of location's class instance and a few instance of person class and organization class. We dealt with sub-ontology: person, location, and organization such as Figure (3).

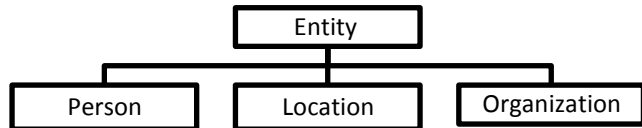


Fig. 3. Sub-ontology of AWN ontology

## VII. DATA EVALUATION

The CONLL evaluation standard metrics of precision, recall and F-measure are used [34]. Precision and recall can be express as shown in the Equation (3) and (4):

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \quad (3)$$

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negativ}} \quad (4)$$

Also the F-measure (F) was used, which is defined as a weighted combination of precision and recall as Equation (5):

$$F - \text{measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

## VIII. EXPERIMENTS

For the evaluation of this system, the two corpora datasets that are mentioned in section (6) are used. Since both datasets do not follow the same tagging conventions as in section (6), training and testing were conducted separately for each dataset. The ANERCorp dataset was used to compare our work with previous works. The datasets is divided into 80% as

<sup>11</sup> <http://users.dsic.upv.es/~ybenajiba/>

<sup>12</sup> [http://www.altec-center.org/Repository\\_65.html](http://www.altec-center.org/Repository_65.html)

<sup>13</sup> <http://users.dsic.upv.es/~ybenajiba/>

<sup>14</sup> [http://en.wikipedia.org/wiki/List\\_of\\_Arabic\\_names](http://en.wikipedia.org/wiki/List_of_Arabic_names)

<sup>15</sup> <http://globalwordnet.org/arabic-wordnet/awn-browser/>

training dataset and 20% as testing dataset according to Abdul-Hamid and Darwish [17], and Kareem Darwish [18].

New Semantic information features have added into CRF. CRF++ tool<sup>16</sup> is used for training and testing. In training step, the CRF++ tool needs two input template files and training data file. The tool output is the classifier model file. The template file describes which features are used in training and testing. In each template, special macro %x[row , col] will be used to specify a token in the input data. Row specifies the relative position of the current focusing token and col specifies the absolute position of the column. In training data file, each word is represented by a set of features and its actual NE's class in order to produce a CRF classifier. In testing step, the tool needs the output of training step (model file) and testing data file. Output of this step is the predication class for each word.

For semantic information features in ontology, the AMN tool is modified to extract all features from ANW ontology for all words in datasets (mapping between the dataset and ANW ontology). The information was extracted that needed, such as person information, location information and organization information. The semantic information features was introduced in a features file as CRF features. Other feature were added called PART, it represents the classes of two words and a relation between them. For example ( القاهرة عاصمة مصر ) [Cairo the capital of Egypt] all sentence is a PART. Contextual window size parameter was used as experimental factor to our feature engineering experiments. Window size significantly effects on NER accuracy. Three type of window size was utilized in this work -1/+1, -2/+2, and -2/+1. Based on the experiments conducted, the window size -2/+1 is best choice for that datasets used in this work.

## IX. RESULTS AND DISCUSSION

The system was trained on the data in cumulative additions of features. That said, the system was trained on first two features (WF and POS), then adding GAZ, and so on. The last added is the semantic information features. Table [1] shows the results of ANERcorp and ALTEC datasets obtained from CRF for all feature sets in terms of precision(P), recall(R) and F-measures(F) for Person, Location and Organization. The best results for P, R, and F are **bolded** in the tables. These results show the effect of using cumulative additions of features on training accuracy. There is the most significant impact on performance when adding GAZ. The second feature is semantic information. When all were combined (Table [1]), the resulting precision is (94.44%) which was almost (0.46%) above the best precision obtained, by WF\_POS features (93.98%). The recall is (82.13%) which was about (1.04%) above the best recall obtained, by WF\_POS\_GAZ\_CF\_GF features (81.09%). In addition, F-measure (87.86%) was most

(0.90%) above the best F-measure achieved, by WF\_POS\_GAZ\_CF\_GF features (86.96%) when used ANERcorp dataset. While when applied on ALTEC dataset, the recall (79.52%) was (0.95%) point over the best recall got, by WF\_POS\_GAZ\_CF\_GF features (78.57%). In addition, F-measure (84.72%) was (1.59%) over the best F-measure acquired, by WF\_POS\_GAZ\_CF\_GF features (83.13%).

The results of the experiment illustrate that the ANER with semantic information (ontology features) can achieve better performance. The precision, recall, and F-measure of semantic information (ontology features) are higher than other features, which means that adding semantic information (ontology) can improve the precision, recall, and F-measure of ANER. The reasons may be that ontology is a kind of concept models that could describe the system at the level of semantics and knowledge.

Tables [2] and [3], summaries the best results of this system on ANERCorp and ALTEC datasets respectively.

In comparison to results with previous work, this system outperforms result of other Arabic NER systems when applied on ANERcorp dataset as shown in Table [4]. It also outperforms the previous systems regarding F-measure in extracting Person, Location and Organization NEs from ANERcorp with an overall F-measure= 87.86 %.

We compare our work with previous works done by Benajiba et al. [4], Abdul-Hamid et al. [17], and Darwish [18], which produce better results than their system with less number of features.

## X. CONCLUSION

This paper presented an integration of features set fort named entity recognition in Arabic. This integration combines internal features that represented linguistic features as well as external features to represent the semantic of Arabic language. The internal features such POS, GAZ, indicator, and Cram character features while the external features is semantic information features were extracted from Arabic wordNet ontology such as classes, instance and relations.

The integration model helped overcome some of the orthographic and morphological complexities of Arabic. Experimental results show F-Measure for ANERCorp and ALTEC around 87.86% and 84.72% respectively. The proposed feature set achieved improved results over those in the literature with as much as 3.56% F-measure improvement for recognizing NE.

In the future, we intend to study the possibility of improving the system performance using other approaches such as Ruled Based approach and Hybrid approach with semantic information features.

<sup>16</sup> <https://taku910.github.io/crfpp/#download>

TABLE I. RESULTS FOR SUCCESSIVE ADDITION OF FEATURES ON ANERCORP AND ALTEC DATASETS

Feature sets	Type	ANERCorp			ALTEC		
		P	R	F	P	R	F
WF + POS	PERS	93.26	58.14	71.63	<b>92.81</b>	61.92	74.28
	LOC	<b>95.35</b>	71.69	81.85	<b>93.27</b>	80.16	86.22
	ORG	<b>93.33</b>	53.85	68.29	<b>89.66</b>	64.11	74.76
	Overall	93.98	61.23	74.15	<b>91.91</b>	68.73	78.65
WF + POS+ GAZ	PERS	95.20	84.25	89.39	89.91	74.68	81.59
	LOC	94.05	84.87	89.22	86.74	85.76	86.25
	ORG	91.32	72.25	80.68	83.23	74.80	78.79
	Overall	93.52	80.46	86.49	86.63	78.41	82.32
WF + POS+ GAZ+ CF	PERS	<b>96.06</b>	81.80	88.36	90.02	75.20	81.95
	LOC	92.57	87.87	90.16	89.43	85.33	87.29
	ORG	92.83	71.15	80.56	85.81	71.23	77.85
	Overall	93.82	80.28	86.52	88.39	77.26	82.45
WF + POS+ GAZ+ CF + GF	PERS	93.90	84.28	88.83	90.11	74.04	81.29
	LOC	94.60	88.85	91.63	90.70	85.78	88.17
	ORG	92.75	70.14	79.88	83.94	<b>75.89</b>	79.71
	Overall	93.75	81.09	86.96	88.25	78.57	83.13
All	PERS	95.44	<b>85.13</b>	<b>89.99</b>	91.21	<b>78.80</b>	<b>84.55</b>
	LOC	94.59	<b>88.94</b>	<b>91.68</b>	92.65	<b>86.90</b>	<b>89.68</b>
	ORG	93.29	<b>72.33</b>	<b>81.48</b>	88.08	72.88	<b>79.76</b>
	Overall	<b>94.44</b>	<b>82.13</b>	<b>87.86</b>	90.65	<b>79.52</b>	<b>84.72</b>

TABLE II. BEST RESULTS ON ANERCORP DATASET

Type	P	R	F
PERS	95.44	85.13	89.99
LOC	94.59	88.94	91.68
ORG	93.29	72.33	81.48
Overall	94.44	82.13	87.86

TABLE III. BEST RESULTS ON ALTEC DATASET

Type	P	R	F
PERS	91.21	78.80	84.55
LOC	92.65	86.90	89.68
ORG	88.08	72.88	79.76
Overall	90.65	79.52	84.72

TABLE IV. COMPARISON WITH OTHER ARABIC NER SYSTEMS ON ANERCORP DATASET

System	Person	Location	organization	Overall
	F-Measure	F-Measure	F-Measure	F-Measure
CRF-based system [4]	73.35	89.74	65.76	79.21
Abdul-Hamid and Darwish [17]	82.00	88.00	73.00	81.00
Kareem Darwish [18]	82.10	90.00	72.90	84.30
<b>Our System</b>	<b>89.99</b>	<b>91.68</b>	<b>81.48</b>	<b>87.86</b>

REFERENCES

[1] K. Shaalan, "A survey of Arabic named entity recognition and classification." *Computational Linguistics* vol. 40, no. 4, pp. 469-510, 2014.

[2] S. AbdelRahman, M. Elarnaoty, M. Magdy, & A. Fahmy, "Integrated machine learning techniques for Arabic named entity recognition." *IJCSI*, vol. 7, pp. 27-36, 2010.

[3] N. F. Mohammed, and N. Omar, "Arabic named entity recognition using artificial neural network." *Journal of Computer Science* vol. 8, no. 8, pp. 1285, (2012).

[4] Y. Benajiba, and P. Rosso, "Arabic named entity recognition using conditional random fields." *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*. vol. 8. 2008.

[5] Y. Benajiba, M. Diab, and P. Rosso, "Arabic named entity recognition using optimized feature sets." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008.

[6] Y. Benajiba, M. Diab, and P. Rosso, "Arabic named entity recognition: A feature-driven study." *Audio, Speech, and Language Processing, IEEE Transactions.*, vol. 17, no. 5, pp. 926-934, 2009.

[7] K. Shaalan, and H. Raza, "Person name entity recognition for Arabic." *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. Association for Computational Linguistics, 2007.

[8] K. Shaalan, and H. Raza, "Arabic named entity recognition from diverse text types." *Advances in Natural Language Processing*. Springer Berlin Heidelberg, pp. 440-451, 2008.

[9] K. Shaalan, and H. Raza, "NERA: Named entity recognition for Arabic." *Journal of the American Society for Information Science and Technology*. vol. 60, no. 8, pp. 1652-1663, 2009.

[10] F. Luo, H. Xiao, and W. Chang, "Product named entity recognition using conditional random fields." *Business Intelligence and Financial Engineering (BIFE), 2011 Fourth International Conference on*. IEEE, 2011.

[11] U. Yasavur, R. Amini, C. L. Lisetti, & R. Naphali, "Ontology-Based Named Entity Recognizer for Behavioral Health." *FLAIRS Conference*. 2013.

[12] D. Sánchez, D. Cisneros, and F. A. Gali, "UEM-UC3M: an ontology-based named entity recognition system for biomedical texts." *Association for Computational Linguistics*, 2013.

[13] M. Aboaga, and M. I. Ab Aziz, "Arabic person names recognition by using a rule based approach." *Journal of Computer Science*. vol. 9, no. 7, pp. 922, 2013.

[14] A. Elsebai, F. Meziane, and F. Z. Belkredim, "A rule based persons names Arabic extraction system." *Communications of the IBIMA*. vol. 11, no. 6, pp. 53-59, 2009.

[15] Y. Benajiba, M. Diab, and P. Rosso, "Arabic named entity recognition using optimized feature sets." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008.

[16] K. Darwish, and W. Gao, "Simple Effective Microblog Named Entity Recognition: Arabic as an Example." 2014.

[17] A. Abdul-Hamid, and K. Darwish, "Simplified feature set for Arabic named entity recognition." *Proceedings of the 2010 Named Entities Workshop*. Association for Computational Linguistics, 2010.

[18] K. Darwish, "Named Entity Recognition using Cross-lingual Resources: Arabic as an Example." *ACL (1)*. 2013.

- [19] A. Zirikly, and M. Diab, "Named Entity Recognition for Dialectal Arabic." *ANLP 2014*, pp. 78 2014.
- [20] Y. Benajiba, M. Diab, and P. Rosso, "Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition." *Int. Arab J. Inf. Technol.* vol. 6, no. 5, pp. 463-471, 2009.
- [21] Y. Benajiba, P. Rosso, and J. M. Benedíruiz, "Anersys: An arabic named entity recognition system based on maximum entropy." *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, pp. 143-153, 2007.
- [22] O. Zayed, S. El-Beltagy, and O. Haggag, "An Approach for Extracting and Disambiguating Arabic Persons' Names Using Clustered Dictionaries and Scored Patterns." *Natural Language Processing and Information Systems*. Springer Berlin Heidelberg, pp. 201-212, 2013.
- [23] A. Hassan, H. Fahmy, and H. Hassan, "Improving named entity translation by exploiting comparable and parallel corpora." *AMML07 2007*.
- [24] D. Samy, A. Moreno, and J. M. Guirao, "A proposal for an Arabic named entity tagger leveraging a parallel corpus." *International Conference RANLP, Borovets, Bulgaria*. 2005.
- [25] E. Kamal, M. Rashwan, and S. Alansary, "High Quality Arabic Lexical Ontology Based on MUHIT, WordNet, SUMO and DBpedia." *Computational Linguistics and Intelligent Text Processing*. Springer International Publishing, pp. 98-111, 2015.
- [26] S. Elkateb, W. Black, and P. Vossen, "Building a wordnet for arabic." *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*. 2006.
- [27] M. Althobaiti, U. Kruschwitz, and M. Poesio, "A Semi-supervised Learning Approach to Arabic Named Entity Recognition." *RANLP*. 2013.
- [28] O. Zayed, S. El-Beltagy, and O. Haggag, "A novel approach for detecting Arabic persons' names using limited resources." *Complementary Proceedings of 14th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing*. 2013.
- [29] M. Althobaiti, U. Kruschwitz, and M. Poesio, "Combining Minimally-supervised Methods for Arabic Named Entity Recognition." *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 243-255, 2015.
- [30] T. T. Nguyen, and T. H. Cao, "Linguistically Motivated and Ontological Features for Vietnamese Named Entity Recognition." *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2012 IEEE RIVF International Conference on*. IEEE, 2012.
- [31] W. Zaghouani, "RENAR: A rule-based Arabic named entity recognition system." *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 11, no. 1, 2012.
- [32] J. Lafferty, A. McCallum, and F. CN. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." 2001.
- [33] I. Boujelben, S. Jamoussi, and A. Hamadou, "A hybrid method for extracting relations between Arabic named entities." *Journal of King Saud University-Computer and Information Sciences*, vol. 26, no. 4, pp. 425-440, 2014.
- [34] A. De Sitter, T. Calders, and W. Daelemans, "A formal framework for evaluation of information extraction." *Online <http://www.cnts.ua.ac.be/Publications/2004/DCD04>*, 2004.

# Clustering Analysis of Wireless Sensor Network Based on Network Coding with Low-Density Parity Check

Maria Hammouti

Dept. Electronics, Informatics and Telecommunications  
ENSAO, Mohammed I University  
Oujda, Morocco

Kamal Ghoumid

Dept. Electronics, Informatics and Telecommunications  
ENSAO, Mohammed I University  
Oujda, Morocco

El Miloud Ar-reyouchi

Dept. Telecommunication and Computer Science  
Abdelmalek Essaadi University  
Tetouan, Morocco

Ahmed Lichioui

Broadcasting TV/FM  
SNRT  
Rabat, Morocco

**Abstract**—The number of nodes in wireless sensor networks (WSNs) is one of the fundamental parameters when it comes to developing an algorithm based on Network Coding (NC) with LDPC (Low Density Parity Check) code because it directly affects the size of the generator matrix of the LDPC code and to its dispersion. Optimizing Wireless Communication Systems by decreasing BER (Bit Error Rate) is one approach to analyze the network into clusters (at the level of their nodes). In this paper, the authors present a fully distributed clustering algorithm and they consider different node values by cluster, then they select the curves that have the best compromise. They examine the effects of SNR (Signal-to-noise ratio) quantization on system performance obtained for different scenarios (by varying the parameter corresponding to the number of the symbol during the forwarding phase). Finally, the results prove that the increased number nodes improve LDPC code properties.

**Keywords**—Clustering Techniques; Network Coding; LDPC codes; distributed algorithms; wireless sensor network

## I. INTRODUCTION

A wireless sensor network (WSN) is a group of specialized transducers with a communications infrastructure for target tracking, environmental monitoring and recording conditions at diverse locations. WSN consists of multiple detection stations called sensor nodes, each of which is small, lightweight and portable.

Hierarchical clustering [1], [2] is particularly useful for applications that require scalability to hundreds or thousands of nodes. Moreover, clustering [3] can stabilize the network topology at the level of sensors and thus cuts on topology maintenance overhead. The need for clustering in WSNs has been first motivated, and a brief description of the implied hierarchical network pattern has been given. The sensor nodes periodically transmit their data to the corresponding particular node called cluster head (CH) nodes. The CH nodes aggregate the data and broadcast them to the base station (BS) either directly or through the intermediate communication with other CH nodes.

NC [4] is a recent field of the information theory that breaks with this assumption. Instead of only forwarding data, nodes may recombine several input packets into one or several output packets.

The concept of NC has been first introduced for satellite communications in [5] and then fully developed in [4] for general networks.

With NC, a network node allows combining several packets (messages) that it has generated or received into one or several outgoing packets. NC replaces the conventional routing (data packets have given by store-and-forward mechanisms in which the intermediate nodes only repeat data packets that they have received).

Recently, NC has been applied to wireless networks and received significant popularity as a means of improving network capacity and the End-to-End Delays of transmissions [6]. It can also increase the capacity and reduces the delay [7] in networks and at point-to-point communications [8], in broadcast networks [9]. NC increases the energy efficiency in broadcast networks [10]; it optimizes capacity and energy consumption in the network [11], [12]. Also, it can be combined with routing for improved performance [13], and adaptive algorithms of appropriate node mobility networks [14].

In fact, the unreliability and broadcast nature of wireless links make wireless networks a natural setting for network coding. Moreover, network protocols in wireless networks, e.g., wireless mesh networks and mobile ad hoc networks, are not fully developed yet and hence there is more freedom to apply network coding in such environments compared to wire line networks such as the Network Management System [6].

Low-Density Parity Check (LDPC) codes [15] deliver very good performance when it decoded with SPA [16]. As LDPC codes are being used in a wide range of applications, the search for efficient implementations of decoding algorithms is being pursued intensively.

Error-correcting coding schemes using LDPC codes and belief propagation decoders based on the sum-product algorithm (SPA) have recently achieved some of the highest performance results in the literature. The SPA has been presented to decode LDPC codes.

The main contribution in this paper is reducing the BER value (through a distributed coding using LDPC codes) in ANCC (Adaptive Coded Cooperation Network) for WSN by using a very effective method adapted to this type of network. It studies two novel scopes in communication. The authors focus on the analysis of the BER. Regarding NC, they propose a coding algorithm adequate for dense and extensive. They create a set of scenarios in which several parameters have been modified to analyze how they affect the BER versus SNR curve. What is more, this paper aims to get the right values of these parameters to find a compromise, which seeks to decrease the BER

The rest of the paper is organized as follows. In Section II, the model on which to base our approach is described. In section III, the authors discuss the results; they compare the different approaches, and they analyze their applicability. The conclusions as well as future work are discussed in Section IV.

## II. NETWORK MODEL OVERVIEW

### A. Description Working Scenario and ANCC Scheme

The network coding scheme to being used as a basis for development is the ANCC [18],[19] that adapt precisely to extensive and dense network like WSN, on which each fixed period, the N sensor nodes should send the information collected to the CH. To avoid interference, in the type of WSN in question, it is often established access scheme TDMA (Time Division Multiple Access) in which each node has a particular time for the transmission while all other nodes remain in silence.

The fundamental idea of the scheme ANCC network coding is that neighboring nodes can linearly combine the correctly received information and send it to the cluster head in a later phase of the communication. In particular, the ANCC scheme uses LDGM codes (Low Density Generator Matrix) [19]. This type of LDPC code fits to WSN scenario by the structure of its generator matrix. In LDGM codes, the generator matrix  $\mathbf{G}$  is being constructed as the concatenation of two matrices: the identity matrix  $\mathbf{I}_N$  and one dispersed and random matrix  $\mathbf{P}$ . The matrices involved in encoding and decoding are as follows [19]:

- Generator matrix  $\mathbf{G}$ , of size  $N \times 2N$  (since the encoding rate is  $1/2$ ; the matrix  $\mathbf{P}$  is of size  $N \times N$ ):

$$\mathbf{G} = [\mathbf{I}_N, \mathbf{P}] \quad (1)$$

- Parity check matrix  $\mathbf{H}$  of size  $N \times 2N$ :

$$\mathbf{H} = [\mathbf{P}^T, \mathbf{I}_N] \quad (2)$$

- Syndrome vector  $\delta$  this vector is the result of checking a word with the parity check matrix  $\mathbf{H}$ . When the decoding performed by syndrome, we must have syndrome validates a coded word  $c$  must be the null vector as follows from:

$$\delta = c.H^T \quad (3)$$

The Network Coding scheme ANCC gets applied the coding LDGM of distributed form:

1) *Diffusion phase (or initial transmission)*: In this phase, each of the N nodes in the WSN transmitted orderly (with TDMA) its corresponding information symbol (systematic symbol) to the head cluster, while the rest of the nodes remain in listening mode.

2) *Forwarding phase (or combined retransmission combined)*: the N nodes transmit a linear combination (coded symbol). It composed of symbols  $\mathbf{s}$ , randomly chosen from the symbols of a node, which had received correctly in the previous phase.

3) *Decoding phase*: once the Forwarding phase has been completed, the CH has already received all the symbols (systematic and codified) and it is capable constructing the parity check matrix  $\mathbf{H}$ , using, for this, the headers received in Forwarding phase. After building the matrix  $\mathbf{H}$ , the cluster head decodes the received symbols obtaining gain codification (decrease of BER) proper to the use of an LDGM code.

### B. Topologies

Hierarchical network allows assigning different roles to nodes; exploiting that to the control node and link activity. Grouping sensor nodes into clusters has been widely pursued by the research community to achieve the network scalability objective. Every cluster would have a leader, often referred to the CH. The strategy is to divide the total set of nodes in several clusters or groupings involving a high number of nodes of each one. To ensure the functioning of the system, each cluster is chosen so that all its elements may have a direct link to the most important node of the cluster, which has already called the cluster head (CH) node (see Figure 1). This node takes charge of being the destination node broadcasts performed by the other nodes of the cluster, being the element that receives and decodes all the information of this grouping (cluster) and, therefore, allows data communication in a clustered network.

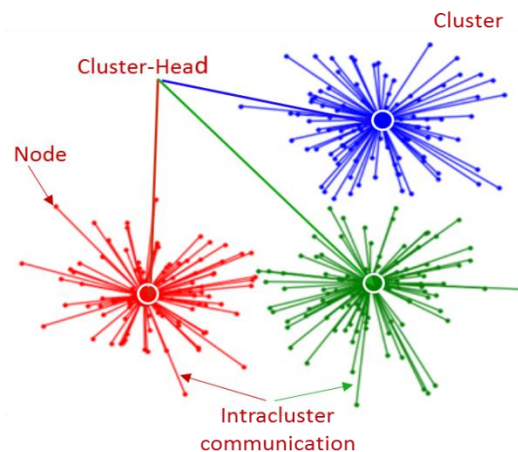


Fig. 1. Network model with clustering



C. Parameters affecting our system

Grouping sensor nodes into clusters has been widely pursued by the research community to achieve the network scalability objective. Every cluster would have a leader often referred to as the CH.

The main parameters to have analyzed for the scenario are:

1) *The number of iterations of the Sum-Product algorithm SPA:* It is analyzed for a concrete example the improvement of having a greater number of iterations in the SPA.

2) *The parameter s:* Several options were tested to see how it affects the number of nodes involved as a response of linear combination sent by each node during forwarding phase to the BER obtained in an ANCC scheme.

3) *The number of nodes per cluster:* In this case, it is mainly analyzed how the number of nodes that intervene in the distributed encoding LDPC affects the BER and how the results present themselves for various options.

We define a relationship between the symbols, which has combined in the forwarding phase and the number of nodes that can intervene in this combination. The Dispersion Grade (DG) refers to the dispersion having the generator matrix of the LDPC code, this DG is calculated as:

$$DG = \frac{s}{N - 1} \tag{4}$$

Where *s* is a parameter that indicates the values of the coded symbol (randomly chosen) which are correctly received in the previous phase, *N* is the total number of nodes. A parameter DG is the percentage of symbols involved in the response of forwarding phase (among the symbols that could intervene). The goal is to see if it can get a range of values of DG between which algorithm behave optimally, with in order to establish a priori the values of the parameter *s* and the number of nodes that give the best BER-SNR performance.

D. Methods of Performance Evaluation

A typical BER curve, for LDPC code, is shown in figure 2. Three regions [20] namely low SNR region, waterfall region, and error floor region, can be identified. In the low SNR region, BER decreases slowly as SNRT increases. For the intermediate value of SNR, the BER decreases rapidly in the waterfall region with an increase in SNR. In this region, the coding gain approaches the theoretical limit. The error Floor is due to the weight distribution of LDPC codes. Normally, LDPC codes do not have large minimum distances. Hence, lowering the Error floor region result in better codes, which in some cases, may result in faster convergence in decoding.

The method have to be followed to determine which scenario is better than another is based on the inspection of their BER versus SNR curves, choosing to get a better compromise between the values of BER obtained at high and low values of SNR. The studies carried out show that there is a relationship between the position of the threshold *th<sub>1</sub>*, the slope in the Waterfall region and BER values obtained in the of Error Floor region.

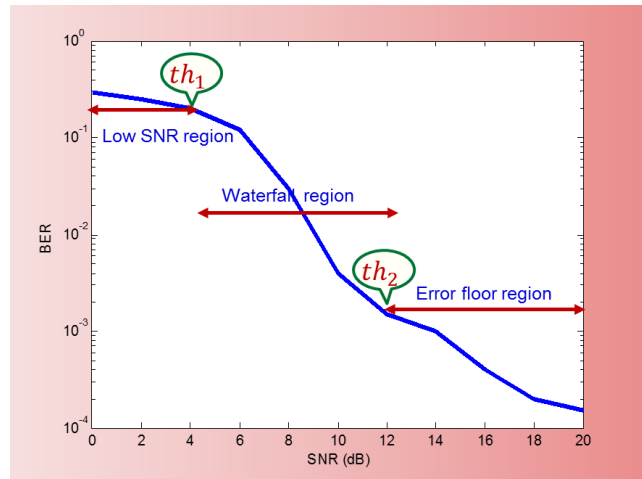


Fig. 2. Regions BER-SNR curve for a LDPC codes

Therefore, the goal is to find a compromise among which this threshold *th<sub>1</sub>*, it is located at the lowest possible, so that the SNR for BER values obtained in the region of Error Floor will be better and this region appears for reasonable SNR values in a wireless communication.

III. RESULTS AND DISCUSSION

In this section, we analyze the data results obtained from simulations for the clustering analysis of different scenarios, which corresponds to the sending and forwarding the data nodes of each aggrupation of their cluster head node. We start by cutting the network in groups of nodes called clusters, thereby giving, at the network, a hierarchical structure [21].

Within each phase, behavior is also studied in each cluster for different values of the parameter *s*.

For the simulations, we use a total of 1500 TDMA network nodes, depending on the number of nodes per cluster and so that it can give an exact value of cluster head nodes. We have 1500 nodes, which means N=1500 nodes.

The number of nodes defined as N, the size of the generator matrix will be N x 2 N. This leads to one of the objectives is to make a cluster of large size, encompassing the largest possible number of nodes. If the total number of nodes is defined as N<sub>Total</sub> and divides into regular cluster of N nodes, therefore, number of cluster head node N<sub>Cluster</sub> is:

$$N_{Cluster} = \frac{N_{Total}}{N} \tag{5}$$

Table I gives the exact value of cluster head.

TABLE I. TOTAL NUMBER (1500 NODES) DIVIDED BY CLUSTER NODE FOR OBTAINED AN EXACT CH NUMBER

Number of nodes	Number of cluster head
150	10
250	6
375	4
500	3



### A. Influence of the number of iterations

Before embarking on an approach, analyzing the results of different nodes per cluster there is a need to explore the influence of iteration number. By Sum-Product Algorithm, we measured, with MATLAB Profiler tool, a distributed LDPC coding for a WSN with 100 nodes.

SPA performs a series of operations on the received symbols so that they can comply with the condition of the syndrome shown in equation 3.

The aim is to analyze if the use of a large number of iterations improve, significantly, BER values obtained to compensate the time delay that occurs during decoding.

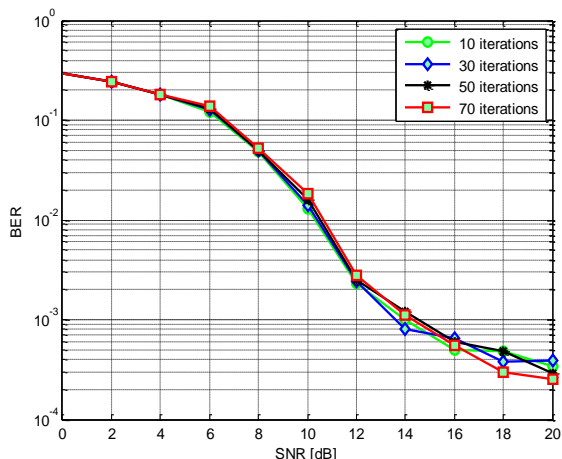


Fig. 3. BER-SNR curve obtained for a distributed LDPC coding in WSN 150 nodes. The combinations of the forwarding phase  $s = 7$  nodes and the maximum number of iterations of the decoding algorithm is varied by taking the values 10, 30, 50 and 70

Figure 3 shows the different curves obtained in simulations for the case of a cluster of 150 nodes where the value of the parameter  $s$  is 7, using different maximum numbers of iterations in the decoding algorithm.

Then we concluded that increasing the number of iterations does not offer a significant advantage. Therefore, from now on, we use 10 (as maximum number for the sum-product algorithm) iterations in our system for all decoders.

### B. Clustering analysis

#### 1) Clusters of 60 nodes

We analyze one grouping of 60 nodes. In this case, our approach is to compare simulation results with experimental data to obtain valid and accurate results.

##### a) Simulation results

With 60 nodes, the generator matrix will have a size of  $60 \times 120$ , which it could be considered a small size; due to high DG. The range in which we study the behavior of the parameter  $s$  is comprised between 4 and 8 in order to offers a better SNR-BER curve.

The figure 4 shows the different curves obtained for a grouping of 60 nodes by varying the value of the parameter  $s$  and using a maximum of 10 iterations of the decoding SPA.

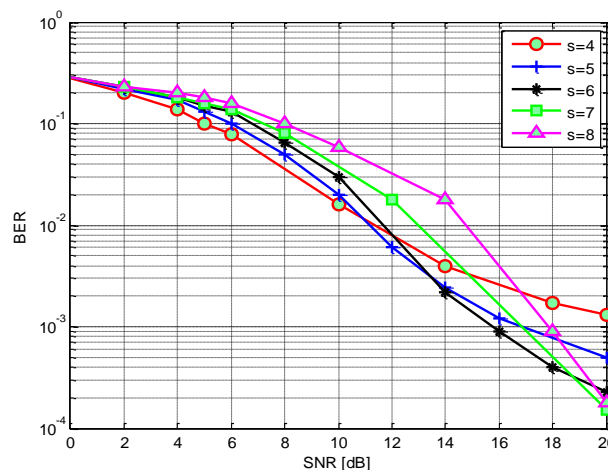


Fig. 4. BER-SNR curve obtained for a distributed LDPC coding in a 60 WSN nodes and the value of the parameter  $s$  varies between 4 and 8

The curve, which presents a better compromise between the position of the threshold values  $u_1$  and BER obtained in the region of error floor, is the curve with  $s = 6$ . DG (with the principal parameters) presenting these curves are summarized in Table II.

TABLE II. DG FOR 60 NODES AND DIFFERENT VALUES OF THE PARAMETER S

Parameters $s$	DG
4	6.77 %
5	8.47 %
6	10.16 %
7	11.86 %
8	13.55 %

These values of the DG are too high and may be associated to sparse matrices that widely dispersed, which does not appear good option to use the groupings of this size.

#### b) Experimental Results

To verify our next simulation results, we analysis the experimental results, we choose 60 sensor nodes within a cluster view the Figure 5.

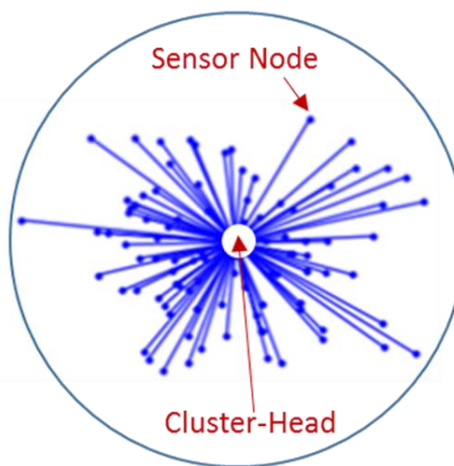


Fig. 5. Experimental topology of the wireless sensor network a (cluster of 60 nodes) the value of the parameter  $s$  is 6

The following figure (Fig 6) shows the experimental results.

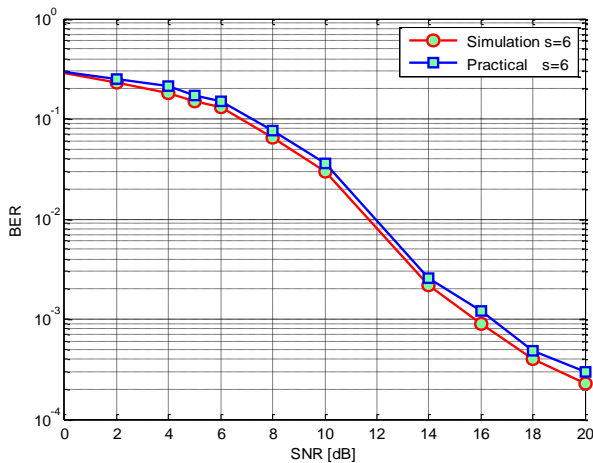


Fig. 6. Comparison between Simulations and Experimental Results obtained for a distributed LDPC coding in a 60 WSN nodes, and the value of the parameter  $s$  is 6

We note that the results of this simulation are very similar indeed; to those obtained by practice. Consequently our preliminary simulation results fully confirm the validity of our next simulations results.

2) Clusters of 150 nodes

With 150 nodes, the generator matrix of network coding with LDPC codes will have a size of 150x300. In addition, the parameter  $s$  is varied from 5 to 11 (to view which of them offers a BER versus SNR curve with a better compromise) while the SPA is kept constant at 10 iterations. Figure 7 shows the different curves obtained for a cluster containing 150 nodes by varying the value of the parameter  $s$ . It can be appreciated that for values of  $s$ , greater than 8, the Error Floor region appears for very high SNR values (over 18 dB), which provokes a good BER obtained at very high SNR values, often situated above the range of SNR in which is located a wireless system.

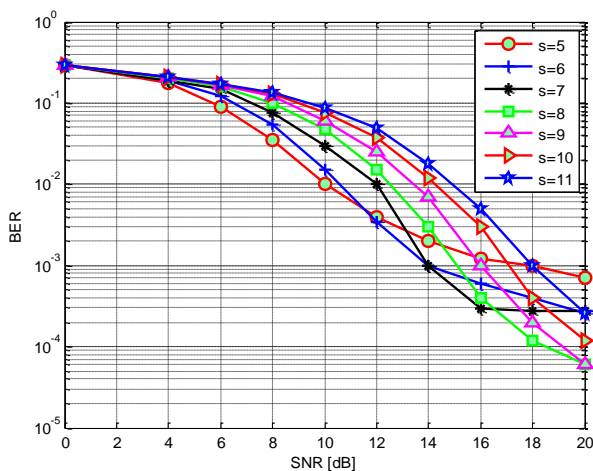


Fig. 7. BER-SNR curve obtained for a distributed LDPC coding in a 150 WSN nodes. The value of the parameter  $s$  varies between 5 and 11

The DG values are summarized in the following table.

TABLE III. DG FOR 150 NODES AND DIFFERENT VALUES OF THE PARAMETER  $S$

Parameters $s$	DG
6	4.02 %
7	4.69 %
8	5.36 %

By Analyzing the graphs, we can see that the curves, with greater compromise, are given for the parameter values  $s= 6, 7$  and 8, whose matrices offer the DG that are displayed in Table III

3) Clusters of 250 nodes

In this case, all nodes are divided into several clusters of 250 nodes. With 250 nodes, the generator matrix of network coding with LDPC codes will have a size of 250x500. In addition, the parameter  $s$  is varied from 7 to 12, while the SPA always keeps the same value 10 iterations.

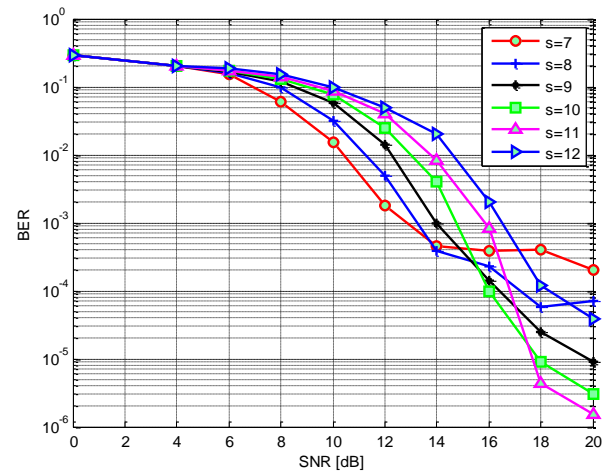


Fig. 8. BER-SNR curve obtained for a distributed LDPC coding in WSN 250 nodes. The parameter  $s$  varies between 7 and 12

The figure 8 represents the different curves obtained for a cluster of 250 nodes by varying the value of the parameter  $s$  and using a maximum of 10 iterations of the decoding SPA.

This figure shows that BER-SNR curves which obtain a greater compromise and whose DG are shown in the Table IV.

TABLE IV. DG FOR 250 NODES AND DIFFERENT VALUES OF THE PARAMETER  $S$

Parameters $s$	DG
8	3.21 %
9	3.61 %
10	4.01 %

4) Clusters of 375 nodes

In this case, we divide the nodes into clusters of 375 nodes. With 375 nodes, the generator matrix of the NC with LDPC code has a size of 375x750. In addition, as in previous cases, the parameter  $s$  is varied from 7 to 11 and BER versus SNR.

The Figure 9 shows the different curves obtained for a cluster of 375 nodes, varying the value of the parameter  $s$ . In

all, the simulations were performed using a maximum of 10 iterations of the algorithm Sum-Product decoding

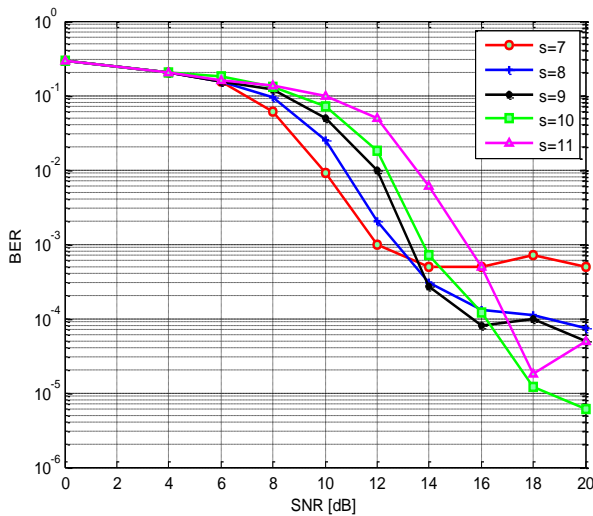


Fig. 9. BER-SNR curve obtained for a distributed LDPC coding in WSN 375 nodes. The parameter  $s$  varies between 7 and 11

By observing the resulting graphics to these groups, it can be seen that the best curves are obtained for values of the parameter  $s$  between 8 and 10. For these parameter values its DG are marked in Table V.

TABLE V. DG FOR 250 NODES AND DIFFERENT VALUES OF THE PARAMETER  $S$

Parameters $s$	DG
8	2.13 %
9	2.40 %
10	2.67 %

### 5) Clusters of 500 nodes

In the final case, the total nodes are divided into several clusters (grouping) so that each cluster includes 500 nodes.

Figure 10 shows the curves obtained in simulations clusters of 500 nodes, using a maximum of 10 iterations of the decoding algorithm Sum-Product

With 500 nodes, the generator matrix of the LDPC code has a size of  $500 \times 1000$ . In addition, the parameter  $s$  is varied between 8 and 13 and the BER simulation results versus SNR are shown in the following figure.

By observing the resulting graphics to these groups, it can be seen that the curves, that offer a better compromise, are given for the parameter values  $s$  comprised between 9 and 12. The DGs obtained are detailed in Table VI.

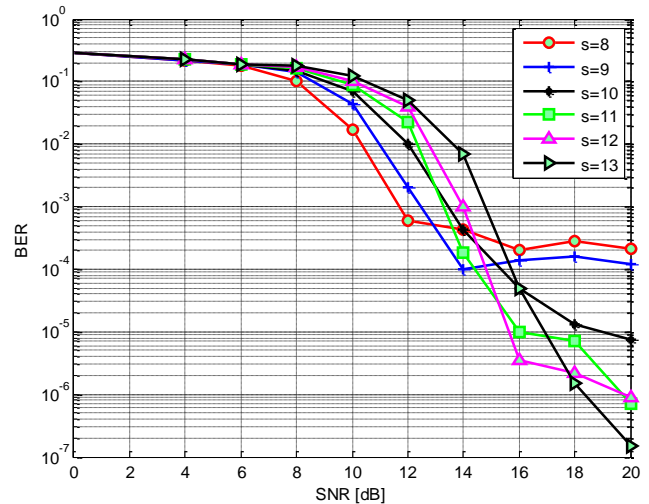


Fig. 10. BER-SNR curve obtained for a distributed LDPC coding in WSN 500 nodes. The parameter  $s$  varies between 8 and 13

TABLE VI. DG FOR 500 NODES AND DIFFERENT VALUES OF THE PARAMETER  $S$

Parameters $s$	DG
9	1.8 %
10	2 %
11	2.2 %
12	2.4 %

### C. Better options of Clustering analysis

We examine the best BER versus SNR curves obtained for the scenarios described above. We choose for each scenario the best compromise solution.

We compare and analyze results of the proposed scenarios:

- Cluster of 150 nodes and parameter  $s = 7$ .
- Cluster of 250 nodes and parameter  $s = 9$ .
- Cluster of 375 nodes and parameter  $s = 10$ .
- Cluster of 500 nodes and parameter  $s = 11$ .

We have the following figure 11

Dispersion grades that correspond to those curves are shown in Table VII.

The curves represented in the figure 11 are the curves that offer better compromise for each cluster of those studied.

By analyzing the curves obtained in Figure 11, we see that the slope of the curve BER-SNR in the Waterfall region increases when using a greater number of nodes per cluster.

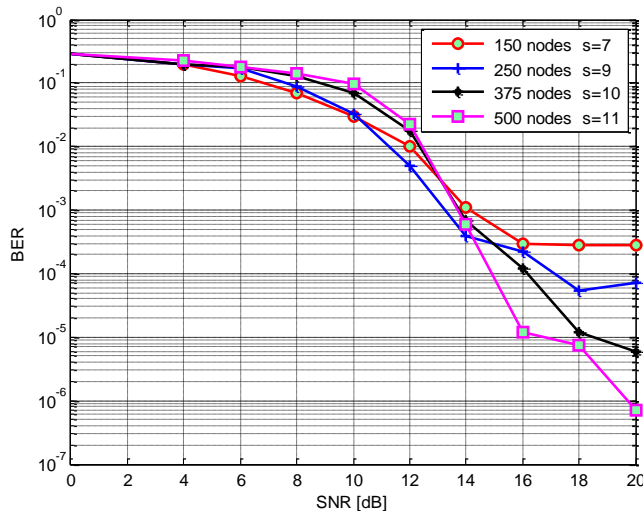


Fig. 11. Comparison of BER-SNR curves obtained for a distributed LDPC coding which offers a better compromise

TABLE VII. DG FOR BETTER OPTIONS OF DIFFERENT VALUES OF THE PARAMETER S

Number of nodes	Parameters <i>s</i>	DG
150	7	4.69 %
250	9	3.61 %
375	10	2.67 %
500	11	2.2 %

#### IV. CONCLUSIONS

We examine and we analyze the best BER curves versus SNR obtained for different scenarios. In light of these results, it can be concluded that the best BER vs. SNR curves obtained correspond to the clusters of 375 and 500 nodes. As a result, we prove that an increasing the number of the node improves LDPC code properties. The application of the LDPC code with NC can improve consistently, considerably and significantly the BER versus SNR when we increase the number of node per cluster in Hierarchical clustering for WSN. On the other side, it is also important to mention that the increase of iterations number has a negligible impact on the performance of our system.

Our future work (which is the next phase of the present paper) is to improve the BER in WSN at the level of clusters heads using NC with low-density parity-check (LDPC) codes.

#### ACKNOWLEDGMENT

This research was made possible in part by the « Société nationale de radiodiffusion et de télévision »SNRT.

#### REFERENCES

[1] C.R. Lin and M. Gerla, Adaptive clustering for mobile wireless networks, *IEEE Journal on Selected Areas Communications*, 15(7), 1997,pp:1265–1275.  
[2] S. Banerjee and S. Khuller, A clustering scheme for hierarchical control

inmulti-hop wireless networks, in *Proceedings of 20th Joint Conference of the IEEE Computer and Communications Societies (INFOCOMS 01)*, Anchorage, AK, April 2001.  
[3] D. Estrin, R. Govindan, J. Heidemann, and S. Kumar, Next century challenges: scalable coordination in sensor networks, in *Proceedings of the ACM/IEEE MOBICOM Intl. Conference*, Boston, MA, August 2000, pp. 6–11.  
[4] R. Ahlswede, N. Cai, S. R. Li, and R. W. Yeung. Network information flow. *IEEE Transactions on Information Theory*, July 2000.  
[5] R. W. Yeung and Z. Zhang, “Distributed source coding for satellite communications,” *IEEE Trans. Inform. Theory*, 45(4), May 1999, pp. 1111–1120.  
[6] El Miloud AR REYOUCHII, Kamal Ghoumid, Koutaiba Ameziane, and Otman El Mrabet, “The Improvement Of End To End Delays In Network Management System Using Network Coding”, *International Journal of Computer Networks & Communications (IJCNC)*, 5(6), Novembre 2013, pp:65-84.  
[7] El Miloud Ar reyouchi, Kamal Ghoumid, Koutaiba Ameziane, and Otman El Mrabet, “The potentials of Network Coding for improvement of Round Trip Time in wireless Narrowband RF communications”, *Publisher: IEEE*, 2014, pp: 765 – 770.  
[8] Katti, S., Rahul, H., H., W., Katabi, D., Medard, M., y Crowcroft, J. Xors in the air: Practical wireless network coding. *IEEE/ACM Transactions on Networking*, 16(3), 2008:pp:497-510.  
[9] Eryilmaz, A., Ozdaglar, A., Medard, M., y Ahmed, E. On the delay and throughput gains of coding in unreliable networks. *IEEE Transactions on Information Theory*, 54(12)2008:pp:5511-5524.  
[10] Fragouli, C., Widmer, J., y Le Boudec, J.-Y. Efficient broadcasting using network coding. *IEEE/ACM Transactions on Networking*, 16(2)2008:450-463.  
[11] Sagduyu, Y. E. y Ephremides, A. On joint mac and network coding in wireless ad hoc networks. *IEEE Transactions on Information Theory*, 53(10), 2007:3697-3713.  
[12] Sagduyu, Y. E. y Ephremides, A. Cross-layer optimization of mac and network coding in wireless queuing tandem networks. *IEEE Transactions on Information Theory*, 54(2)2008:554-571.  
[13] Sengupta, S., Rayanchu, S., y Banerjee, S. An analysis of wireless network coding for unicast sessions: The case for coding-aware routing. In *Proceedings of the IEEE International Conference on Computer Communications*, 2007. *INFOCOM’07,2007*, pp. 1028-1036.  
[14] Yang, D.-N. y Chen, M.-S. Data broadcast with adaptive network coding in heterogeneous wireless networks. *IEEE Transactions on Mobile Computing*, 8(1)2009:109-125.  
[15] R. G. Gallager, *Low Density Parity Check Codes*. Cambridge, MA: MIT Press, 1963.  
[16] D. J. C. MacKay, “Good error-correcting codes based on very sparse matrices,” *IEEE Trans. Inf. Theory*, vol. 45, no. 3,1999: pp. 399–431.  
[17] Xingkai Bao and Jing Li. Matching code-on-graph with network-on-graph: Adaptive network coding for wireless relay networks. In *Proc. Allerton Conf. on Commun., Control and Computing IL*, 2005.  
[18] Xingkai Bao and Jing Li. Adaptive network coded cooperation (ancc) for wireless relay networks: matching code-on-graph with network-on-graph. *Wireless Communications, IEEE Transactions on*, 7(2) February 2008:574 –583.  
[19] Garcia-Frias, J. y Zhong, W. Approaching shannon performance by iterative decoding of linear codes with low-density generator matrix. *IEEE Communication Letters*, 7(6) 2003 :266-268.  
[20] Tao Tian, C. Jones, J.D. Villasenor, and R.D. Wesel. Construction of irregular ldpc codes with low error floors. In *Communications, 2003. ICC ’03. IEEE International Conference on*, volume 5, may 2003.  
[21] Colette Johnen and Le Nguyen. Self-stabilizing weight-based clustering algorithm for ad hoc sensor networks. In *Algorithmic Aspects of Wireless Sensor Networks,2006*,pp 83–94.

# Multi-Agent Based Model for Web Service Composition

Karima Belmabrouk, Fatima Bendella  
Department of Computer Science  
University of Science and Technology of Oran  
Oran, Algeria

Maroua Bouzid  
GREYC- Campus 2 - Sciences3  
University of Caen, UNICAEN  
Caen, France

**Abstract**—The evolution of the Internet and the competitiveness among companies were factors in the explosion of Web services. Web services are applications available on the Internet each performing a particular task. Web users often need to call different services to achieve a more complex task that can't be satisfied by a simple service. And users often prefer to have the best services responding to their requests. In this context, we should measure the Quality of Service (QoS) which is a very important aspect of Web services in order to offer to the user the best services.

“How can we ensure the composition of different services to respond the user request” is the first problem that we contribute to resolve, proposing a multi-agent based model for the automatic planification of Web services. And “guarantee the required quality of the composite Web services” is a complex task regarding the unpredictable nature and dynamics of composite Web services, so our contribution to remedy to this problem consists of the use of two classes of quality attributes. The first one considers generic and the second contains specific attributes.

**Keywords**—Agents; Model; Quality of Service; Service composition; Web services

## I. INTRODUCTION

Web services are applications available on the Internet performing each a special task, such as booking a flight ticket. However, the use of Web services is very limited because they perform only one special task; for example, to go on vacation, a person will find a Web service for each booking he wants to achieve (plane, train, and hotel).

To resolve this problem, Web services can be combined manually to form a single service; it is the composition of Web services.

There are several standards used specifically to describe, locate and compose Web services. We can mention for example: UDDI for Universal Description, Discovery and Integration, WSDL for Web Services Description Language, WSCI for Web Service Choreography Interface, WSFL for Web Service Flow Language and more recently WS-BPEL for Web Services Business Process Execution Language [3, 4, 9, 12 and 13].

The UDDI provides the discovery of services. In practice, it contains technical information and information on business processes such as the address for accessing Web services, but also much more contextual information such as the name of the responsible for their management, a brief description of their

functionality or the name and the company's business branch on which they depend.

The present article consists of five sections; after a brief introduction, the second section is devoted to the presentation of the problem of the Web services composition. The third one concern exhibition of some research works made to remedy various problems related to the composition of Web services. We move then to the fourth section where we present the concept of quality, and some works based on the quality of Web services, issues in Web services composition.

In the fifth section, we present our multi-agent based model, specifying the detailed functional architecture of each agent proposed in the model.

The last part of this paper presents a conclusion about our proposed model, and we exhibit some perspectives of our future research works.

## II. THE WEB SERVICES COMPOSITION

Web service composition refers to the process of creating a composite service, offering new functionality, from simple existing Web services. Several kinds of process ensure this operation like the dynamic process of discovery, the process of integration and the execution process of these services in a particular order to satisfy a definite need [7].

The composition of Web services has the objective of determining a combination of services based on a client or a user request. On this side, this operation appears as a single service and it will be transparent to the user, even if it represents the combination of several Web services.

## III. STATE OF THE ART OF WEB SERVICES COMPOSITION

Various studies have been made to compose Web services like those presented in [1, 3, 4, 5, 7, 9, and 10]. A planning process can be used to find the correct order of Web services in the composition. As the number of services can be relatively important, it is necessary to add an optimization phase whose purpose is to provide the user with the best service compositions according to special criteria.

The proposed approach of A. Yachir et al. in [11] aims to build a plan for a composition of services after verifying its feasibility.

M. El Falou used different planning techniques to overcome some problems related to the composition of services in [5]. He proposed a distributed agent-based model,



developed architecture to respond to user requests, diagnosed online the status of implementation of the plan and repaired automatically his proposition in the case of fault detection, during the execution.

D. Pellier and H. Fiorino proposed in [9] an original architecture of automatic composition of Web services by planning techniques. They presented a fully distributed planning model in which agents were reasoning together to achieve a common goal predefined by the user, creating a global plan representing a possible composition of their services.

The objective of D. B. Claro [3 and 4] was to compose automatically and optimally Web services dedicated to produce cost estimates using planning services to detect which belong to the composition. She proposed a framework made up of several phases that can negotiate services to automatically provide users with optimum results with generic optimization criteria.

However, users can have their evaluation criteria or even contradict the generic one, which has been the work of P. Albers and O. Licchelli in [1]. They utilized the SPOC prototype presented in [3 and 4] and they used the user profiles in order to improve the performance of this system, and they presented the most adequate solutions to a particular user. Researchers validated this approach using an example of building construction based on the ontology of French public market.

HSN is a model of composition that permits correcting and validating a service made before his execution proposed in [2]. The editor allows the progressive-composition followed by verification of time constraints. That means that the user could locate and correct any temporal conflicts in its specification and he can export the RSH file for future use.

The work presented by N. Temglit et al. in [10] proposed a model for the composition of semantic Web services. In this model the concepts handled by Web services application domain are represented semantically, i.e., operations and static concepts used to describe the properties of Web services.

The convergence between Web services and Semantic Web aims to increase the expressiveness of descriptions and make more efficient management, discovery, composition and invocation of services through a communication protocol and a directory service UDDI (Universal Description, Discovery, and Integration). This protocol allows a "supplier" to register his service and a "consumer" to find the right service.

Researchers proposed different levels of abstraction to the "operation" concept to allow a progressive-access to concrete-services. Thus, two levels of composition at different granularities (abstract and concrete) may be generated which will reuse plans already constructed to meet similar needs, even with modified preferences.

#### IV. SOME WORK ON THE QUALITY OF SERVICE

The term QoS refers to the ability to provide a service (such as a communication medium) according to the requirements of response time and bandwidth.

Many types of research corresponding to the quality of Web services have been published in recent years [14, 15, 16 and 17]. They revolve around different areas of research such as:

- How to define a variety of factors of service quality?
- How to specify QoS runtime information based on their volatility and complexity?
- How to match user needs with existing services regarding quality?
- How to set user preferences for Web services?
- How to perform the classification of similar services on user preferences?
- What are the factors that predict the quality of service in some environmental conditions that involve dependencies between QoS parameters and relationships with contextual factors?

N. Kokash and V. D'Andrea proposed a method based on risks to the assessment of Quality of Service [6]. Their approach allows a simple choice of Web services with consideration of several service quality factors. They studied the risk factors to measure the impact of failures of atomic services in the composition of services. The major disadvantage of this metric is that different compositions require redundant recalculate risks, making the computational approach less effective than methods based on the evaluation of QoS for well-defined attributes.

A new approach, to monitoring the composition of services, has also been proposed by A. Yachir et al. [11], in addition to that of the composition of services, to take into account the dynamic and stochastic aspects of the ubiquitous environment. This second approach takes into account the context of use and the quality of services available. Three mechanisms are the basis of the proposed method, namely the Bayesian learning for the invocation of services, automatic discovery of services and the automatic and dynamic reorganization plan composition of services. Authors of these approaches tested them in various application scenarios. The results have shown their feasibility and their usefulness in robotics ubiquitous.

SPOC, which is a system proposed by D. Claro in [3 and 4], consists of four phases:

- Discovery service that helps to identify services from an ontology UDDIO services;
- Planning that allows you to organize the tasks to be performed;
- Execution that sends a request to each service to get information such as cost, duration, etc.; and
- Optimization whose goal is to provide the user with some compositions optimized according to predefined quality criteria such as cost, price, revenue and reputation, and that using the genetic algorithm NSGA-II.

The proposed approach in [8] is to change, dynamically, the path of execution of a composite Web service when it is necessary to meet the QoS requirements. A prototype has been proposed to evaluate the proposed QoS monitoring and adaptation architecture on all for minimizing the response time of composite Web services.

L. Zeng et al. proposed AgFlow, a prototype of Web services composition with the support of quality characterized by [12]:

- A model of service quality to assess the overall quality of Web services.
- Two alternative approaches to service selection for the execution of composite services.

AgFlow is a platform that provides tools for the definition of ontology services; the specification of composite services using state diagrams; and the connection of the service tasks corresponding to a composite service.

AgFlow platform was used to validate the feasibility and benefit of the proposed approaches. In particular, the greatest composite services incorporating a big number of candidate services components have been created and used to conduct experiments with very encouraging results: the computational cost of the planning phase of a composite service with 80 tasks is to about 8 seconds in a static environment and about 190 seconds in a dynamic environment. The comprehensive planning approach leads to a better quality of service, particularly to reduce costs and execution times.

L. Li et al. Presented in [14] a selection approach based on multi-dimension QoS which introduced multi-dimension QoS to describe the QoS properties of Web service composition and computes the QoS of every dimension of Web service composition. The most important result according to this research is the ability to extract the Web services composition which can best respond to the non-functional constraints from the Web service compositions that can satisfy functional requirement of users.

C. Jatoth and G.R. Gangadharan proposed in [15] two fitness metrics using. First, they choose the best fit services for composition by their local fitness value based on the combination of various QoS factors and prioritization. Then, they use a global fitness value to minimize the computational complexity, around service composition. The experiment results indicated the importance of the proposed approach towards scalable and robust QoS-aware service composition. They also plan to focus their future work on another approach based on fuzzy logic in order to evaluate the local fitness value metric.

Z. Zheng et al. conducted evaluations on user-observed QoS of Web services from distributed locations. Service users invoked great number of Web services under heterogeneous environments on real-world Web services. They presented comprehensive experimental results, and released reusable data sets [16].

W. Khowfa et al. used A Multi-Criteria Decision Making (MCDM) and Quality of Service (QoS) attributes in [17] to

obtain the appropriate non-functional service from similar service's pool available in the cloud.

## V. OUR MULTI-AGENT ARCHITECTURE FOR PLANNING THE COMPOSITION OF WEB SERVICES

Our solution to remedy to the problem of user satisfaction constraints, when their requests need composition of different Web services, is to propose a fully distributed model in which various kinds of agents complete their tasks, jointly, to achieve a common goal preset by the user. Those agents create a global plan to define a possible composition of their services. We will describe our detailed proposition in the following sections.

### A. Basic work

We can represent the set of available Web services by a graph whose nodes represent the input and output attributes as well as Web services. And the arcs between these nodes correspond to be links between input attributes and a Web service or connections between a Web service and its output attributes i.e. the results produced after the execution of the Web service. A Web services composition corresponds to a sub-graph of the graph of available Web services.

M. El Falou proposed two algorithms, in this context [5]. The basic idea is to extract the best locally with the use of a global heuristic based on a local evaluation of the best local plan to achieve the goal, and also on an estimation of the globally distributed heuristic that guides the agent to choose its best local plan, ensuring the completeness and the optimality of the algorithm.

To calculate the global heuristic, each service which joins the group of Web services agents must calculate any links between its local services as well as all relationships with all other services.

An agent calculates the global heuristic associated with a service once for each query, back chaining the goal state to the initial state. The global heuristic always involves estimating a distributed distance of a local state  $q_i^k$  of the agent  $A_i$  to the goal state, taking into consideration the best local plans of the others.

Initially, agent  $A_i$  calculates, locally, the distance from  $q_i^k$  to the nearest node to the goal node. This distance ( $q_i^k, q_i^{k'}$ ), where  $q_i^{k'} = \pi_i[q_i^k]$  is the result of execution best local partial plan  $\pi_i$  at the local state  $q_i^k$ . Then, each other agent  $A_i$  can calculate its local distances distance ( $q_{\alpha_j}^l, q_{\alpha_j}^{l'}$ ) between all pairs of states  $q_{\alpha_j}^l$  and  $q_{\alpha_j}^{l'}$  of all the local states.

We ensure the propagation of these local distances, from the local state  $q_i^k$  to neighboring agents, by using the dependencies between the agents i.e. a link appears between a service of an agent  $A_{\alpha_j}$  and a service of another agent  $A_{\alpha_{j+1}}$ , when the first agent produces effects that are pre-conditions of the last one. And this is to obtain a global heuristic measure that can be defined by:

$$\text{distance}(q_i^k, but) = \text{distance}(q_i^k, q_i^{k'}) + \sum_{\alpha_j \in I/i} \text{distance}(q_{\alpha_j}^l, q_{\alpha_j}^{l'}) \quad \text{where:}$$



$q_{\alpha_j}^l \equiv q_{\alpha_{j+1}}^l$ . i.e. the final state of the  $A_{\alpha_j}$  has a direct external link to the original state of the agent  $A_{\alpha_{j+1}}$ ;

I: the index set of agents  $\{1 \dots n\}$ .

The best local plan  $\pi_i$  of an agent  $A_i$  is one that reaches a state  $q_i^{k'}$  such that:

$$\pi_i[q_i^k] = q_i^{k'} \text{ and } q_i^{k'} = \arg \min \text{Distance}(q_i^k, \text{but}).$$

Upon receipt of a query  $R = (\text{init}, \text{goal})$  by an agent, and once the global heuristic calculated, the extraction of the plan is achieved by a forward strategy from the initial state to the goal state.

At initialization, the system proceeds to color all executable services in the init state.

Then, for each neighbor service  $Sn$  of a colored service  $Sc$ ,  $Sn$  is also colored if and only if all its incoming arcs out colored services. By doing this, we get the executable sub-graph of services; where each service in this graph is an accessible and executable service.

To response to user's exigencies of composite Web services, the system must consider all specified features, to extract the mains criteria to which they must respond.

Such a system should consider the fact that the QoS may have multiple dimensions, and that the QoS of the composite services is, essentially, determined in function of QoS of all services used to compose the global Web service.

For example, the user may be required to minimize execution time while satisfying some constraints regarding price and reliability, while another user may give more importance to price at runtime. QoS approach to service composition is, therefore, essential that maximizes the QoS performances of composite services, taking into account the constraints and preferences set by the user.

UDDI directory can locate on the network the desired Web service. This step is very important because it allows access to directories of potential users of Web services.

### B. Our contribution

We were based on the various descriptions of qualification present in the UDDI for selecting the best Web services that may belong to a composite service. This task will be assigned to each agent when choosing its best locally, taking into account the different qualifiers specified by the client, when running the application.

#### 1) Formal definition of Web services

E is an environment consisting of the following:

SW: is a set of Web services  $SW_i$ ,

$$SW = \{SW_1, SW_2 \dots SW_n\} = \{SW_i / i = 1 \dots n\}.$$

$SW_i$ : is a Web service described by a set of input parameters  $Pe_i$  and output parameters  $Psi$ ,

$$SW_i = (Pe_i, Psi).$$

$Pe_i$ : is a set of input parameters of the Web service  $SW_i$ ,

$Pe_i = \{pe_{i1}, pe_{i2} \dots pe_{im}\}$  such that  $m$  is the number of input parameters ( $m = \text{Card}(Pe_i)$ ),

$$Pe_i = \{pe_{i1}, pe_{i2} \dots pe_{im}\} = \{pe_{ij} / j = 1 \dots m\}.$$

$Psi$ : is a set of output parameters of the Web service  $SW_i$ ,

$Psi = \{psi_1, psi_2 \dots psi_k\}$  such that  $k$  is the number of output parameters ( $k = \text{Card}(Psi)$ ),

$$Psi = \{psi_1, psi_2 \dots psi_k\} = \{ps_{ij} / j = 1 \dots k\}.$$

#### 2) Formal definition of semantic links

LS: is a set of semantic connections between Web services. We can represent LS by a square matrix  $SW \times SW$ .

$$LS = \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}$$

$LS_{ij}$ : is a set of semantic connections between Web services  $SW_i$  and  $SW_j$  defined as follows:

$ls(SW_i, SW_j)$

$$= \begin{cases} 1 & \text{if and only if } \exists ps_{it} \in Psi \text{ and } ps_{jt} \in Pe_j \\ 0 & \text{if } \forall pe_{jt} \in Pe_j, \text{ and } \forall ps_{it}' \in Psi \rightarrow pe_{jt} \neq ps_{it}' \end{cases}$$

$f_{lsij}$ : a function that measures the intensity of the semantic link between two Web services  $SW_i$  and  $SW_j$ , defined as follows:

$$f_{lsij} = \frac{\text{Card}(Psi')}{\text{Card}(Pe_j)} / Psi' = \{ps_{it} \in Psi \text{ and } \forall t, ps_{it} \in Pe_j\}$$

Where:

$f_{lsij} = 1$  if and only if  $\text{Card}(Psi') = \text{Card}(Psi)$  and

$\text{Card}(Psi) = \text{Card}(Pe_j)$ .

This equation is evident, if all output parameters  $ps_{ik}$  of a Web service  $SW_i$  match exactly with all input parameters  $pe_{jt}$  of another Web Service  $SW_j$ .

$f_{lsij} = 0$  if and only if  $\text{Card}(Psi') = 0$ .

That means that we can't find any output parameter  $ps_{ik}$  of a Web service  $SW_i$  in  $Pe_j$ , the set of input parameters of a Web service  $SW_j$ .

#### 3) Formal definition of the criteria of the quality of service

In the same environment defined above, we can integrate the following:

- C: is a set of vectors quality parameters  $C_i$  corresponding to different Web services  $SW_i$ .
- $C_i$ : is a vector of quality parameters,  $C_i = (ci_1, ci_2 \dots ci_k)$  Such that  $k$  is the dimension of vector  $C_i$ .

Relative to a given domain  $D$ , it is possible to classify the attributes of service quality by generic or specific attributes. Such as generic attributes are independent of domains to which they belong and the specific attributes are related to the Web services associated with domain to which they depend on.

The model we propose is composed of two classes of quality attributes:

- CGEN: is the set of generic attributes

- CSPE: is the class of specific attributes

In the case of generic attributes CGEN, we can distinguish between measurable attributes and non-measurable ones. So we extend the notation of those parameters as follows:

- $c_i^m$ : is a measurable attribute of a generic Web service SWi.
- $c_i^{nm}$ : is a generic non measurable parameter of a Web service SWi.

Therefore,  $c_{ij}^m$  the jth measurable generic attribute for Web service SWi and  $c_{ik}^{nm}$  the kth non measurable generic one for the same Web service SWi.

We note that most of the measurable attributes are described using performance parameters like:

- The flow: representing the number of service requests during a time interval.
- The response time: indicates the time required to complete a request for a Web service.
- Reliability: this reflects the ability of a service to perform his duties properly.
- Scalability: representing the service's ability to handle the largest number of operations or transactions during a given period, while maintaining the same performance.
- Robustness: this is the probability that a service can respond correctly to invalid input messages, incomplete ones or with conflicts.
- Availability: the possibility of availability of a service.

There are QoS attributes that are not measurable, but are useful for Web services such as:

- The cost or price of performance: it is the price payable by a Web service client for service.
- The reputation is a measure of the credibility of the service. It depends on the user experience.
- The security is a grouping of a set of qualities including confidentiality, message encryption, and access control.

The question that we can ask about our model is the following: How can we measure a non-measurable generic attribute  $c_i^{nm}$ ?

Within the class of specific attributes CSPE, it is clear that since these kinds of parameters can only be set following the domain belonging of a SWi Web service, then it is impossible to enumerate or classify these elements with the same way. Therefore, any particular parameter, in our model can be represented as follows:

- $c_i^s$  : a specific attribute on a Web service SWi.

#### 4) Description of our model:

##### a) General presentation

We propose an automatic architecture for Web services composition based on agents, in order to achieve a distributed planning leading to the creation of a composite Web service seen as a plan defining precedence relationships and causality between basic services and ensuring number of conditions.

Our model is to make a client request, covering a set of parameters and requiring a certain number of constraints that will be considered by an intelligent agent of cognitive type, named «UserAgent», and which will be responsible for managing the preferences of the user, who form his beliefs. He will be responsible also for the synthesis of these constraints, in order to classify them by type.

Another agent used in our model is the «ServiceAgent». A set of them will be, automatically, generated for the purpose of search services that can respond to the request of the user, to filter these services and ordering them according to their relevance by matching each service implicated with an autonomous agent capable of discovering the various offers related to the requested service to choose the best offer and to invoke the service, if necessary.

The third agent in our architecture is «PlanAgent». It's responsible for the Web services composition involved in our process, trying to propose an optimal planning, in terms of time of execution of the overall plan and taking into account the specific constraints imposed by the client.

Each «ServiceAgent» will collaborate with «UserAgent» and «PlanAgent» to propose a global plan to response to the user query, as shown in Fig. 1.

Collaboration between several «ServiceAgent» can take place to try to provide partial plans corresponding to a part of the user query, representing a response to a partial composite service. In this case, it's possible to destroy all «ServiceAgent» implicated in the creation of the partial plan provided by fusion of local solutions and the birth of a new «ServiceAgent» with assigning roles of all agents destroyed. This agent will be able to collaborate with «UserAgent» and «PlanAgent», to search the optimal global plan.

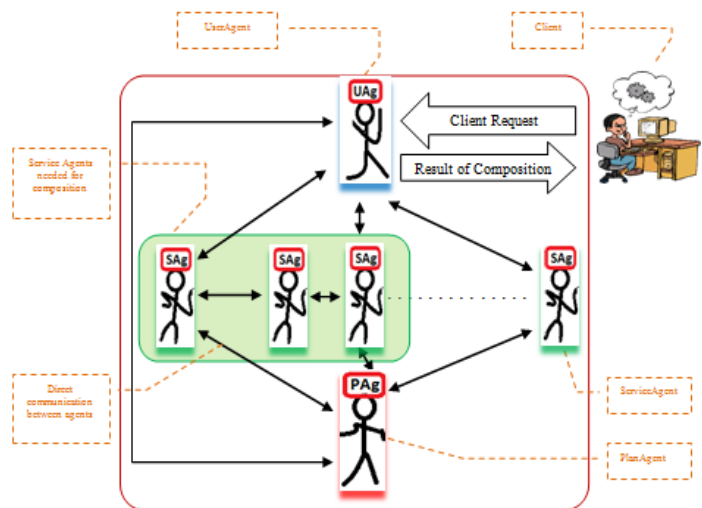


Fig. 1. Agent-based Architecture for Automatic Web services composition

b) The different agent's functional architecture

THE «USERAGENT» ARCHITECTURE:

The «UserAgent» is responsible for processing the user request before forwarding it to the «PlanAgent» will try to find a solution to it. Its primary role is to analyze the application to derive all the parameters related to the purpose and the list of constraints imposed by the user. The «UserAgent» consists of two main modules (see Fig. 2).

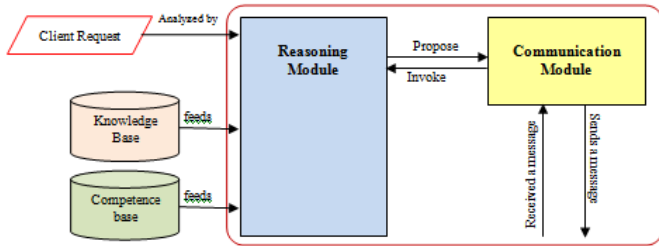


Fig. 2. The «UserAgent» functional architecture

a) The reasoning module: it is the party responsible for the analysis of the user query to determine the parameter list and the input constraints. Based on all knowledge acquired by the agent and his skills, he is able to reason, and to extract the intensions and desires of the user.

b) The communication module: it allows you to manage the communication and interaction with other agents «UserAgent».

THE «PLANAGENT» ARCHITECTURE:

The «PlanAgent» must be able to reason about the task he has to perform and be able to communicate with other agents («UserAgent», «ServiceAgent»). It is initialized with the semantic description of the user to achieve the goal. Its role is to coordinate the work of different agents «ServiceAgent» type to reach the goal.

The «PlanAgent» consists of four modules (see Fig. 3).

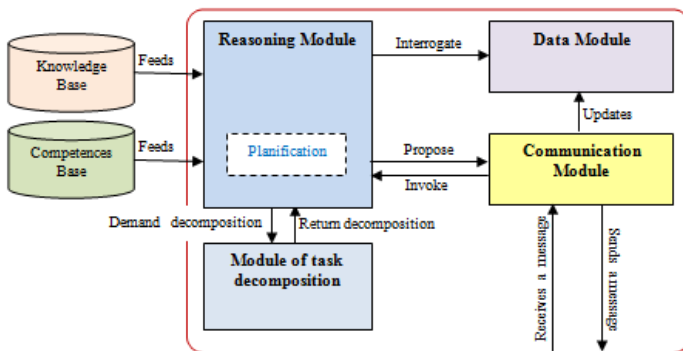


Fig. 3. The «PlanAgent» functional architecture

a) reasoning module: it is the party responsible for defining the behavior of the agent on its interactions with other agents and with the environment, consisting primarily of three essential elements which are:

- The data transmitted by the user, through its request or by other agents.

- Own and skills.
- Knowledge.

b) The data module: which supports the information provided by other agents?

c) The decomposition task module: useful for the decomposition of a major task in elementary sub-tasks where the «PlanAgent» doesn't find a solution to the problem posed by the client.

d) The communication module: it manages the communication and interaction with other agents, and ensures the exchange of messages between them.

THE «SERVICEAGENT» ARCHITECTURE

When the «PlanAgent» cannot find a solution to the query raised by «UserAgent», it invokes The «ServiceAgent. The «PlanAgent», therefore, proceed to the decomposition of the task he could not accomplish to subtasks, and ask agents like «ServiceAgent» to find partial solutions to various elementary tasks. The main role of a «ServiceAgent» is to find a partial plan that meets the client's request, taking into account constraints. It's possible to have talks with a «UserAgent» to negotiate on its preferences. Two essential modules can describe its architecture (see Fig. 4).

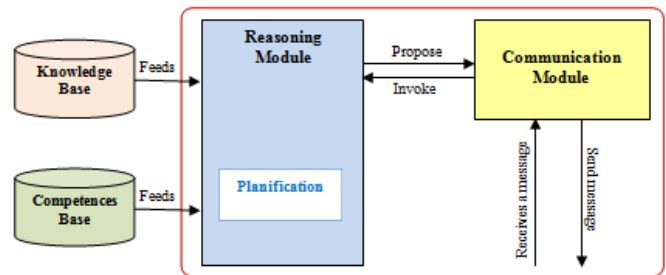


Fig. 4. The «ServiceAgent» functional architecture

a) The reasoning module: it is the party responsible for the search for a partial solution proposed by «PlanAgent» order to contribute to solving part of the problem. The reasoning will be based on the knowledge and skills «ServiceAgent».

b) The communication module: it allows you to manage communication and interaction «ServiceAgent» with another agent which can be either:

- «UserAgent» for the negotiations on the constraints posed.
- Another «ServiceAgent» to collaborate with him in order to get closer to the overall plan, merging their partial plans.
- The «PlanAgent» to provide a partial solution on the stain that has been assigned.

VI. CONCLUSION AND PERSPECTIVES

In this paper, we proposed a multi-agent based model to solve the problem of Web service composition. Our contribution consists of the description of the detailed

functional architecture of each kind of agent used in our prototype.

We suggest to clients to employ two kinds of constraints during the execution of their requests, thus ensuring a better quality of service. A formal presentation of their criteria was introduced to facilitate the evaluation of the optimal plan either in the case of a partial plan or in the case of global one. To validate our model, simulation work is in progress under the «NetLogo» platform where we try to decompose a complex planning problem "P" into N elementary sub-problems "Pi" to find the optimal solution.

An agent type «PlanAgent» is responsible for:

- 1) *Analysis of the query of the «UserAgent»*
- 2) *The search of a «ServiceAgent» able to find a solution to the query Q.*
- 3) *Presentation of the optimal solution found.*
- 4) *If we can't find, the «PlanAgent» is responsible for decomposing the problem Q, into a set of problems Q<sub>j</sub> that we can resolve, separately, by «ServiceAgent»*
- 5) *«PlanAgent» will then be responsible for the composition of partial solutions proposed by the various «ServiceAgent» involved, answering the initial query Q where different strategies in the field of artificial intelligence may be adopted, in this case, among other scheduling algorithms.*

Our simulation environment can contain a set of N agents of type «ServiceAgent», as each of them will be responsible for seeking a global or partial solution to the problem Q.

#### REFERENCES

- [1] P. Albers, O. Licchelli. 2007. «Composition de Services Web Personnalisés», Intelligence Artificielle et Web Intelligence, Grenoble.
- [2] A. Belkheir, S. Bouyakoub. 2009. «A Hierarchical Model for Web Services Composition», International Journal of Web Services Practices, Vol. 4 No.1 (2009), pp.44-50.
- [3] D. B. Claro. 2006. «SPOC—Un canevas pour la composition automatique de services Web dédiés à la réalisation de devis», Thèse de doctorat soutenue le 6 Octobre 2006, Spécialité Informatique; Ecole Doctorale d'Angers, N°d'ordre 767, Université d'Angers.
- [4] D. B. Claro, P. Albers, J. Hao. 2006. «Web Services composition», in J. Cardoso, A. Sheth, Semantic Web Services, Processes and Applications, Chapter 8, Springer, p. 195-225.
- [5] M. El Falou. 2010. «Contributions à la composition dynamique de services fondée sur des techniques de planification et diagnostic multi-agents», Thèse de doctorat soutenue le 2 juin 2010, Université de Caen/Basse-Normandie, U.F.R. Sciences, Ecole doctorale SIMEM.
- [6] N. Kokash and V. D'Andrea. 2006. «Evaluating Quality of Web Services: A Risk-driven Approach»; Technical Report#DIT-06-099.
- [7] B. Medjahed, A. Bouguettaya, and A. K. Elmagarmid. 2003. «Composing Web services on the Semantic Web». The VLDB Journal, 12(4).
- [8] M. Qiao, F. Khendek, A. Serhani, R. Dssouli and R. Glitho. 2009. «An Architecture for Automatic QoS Adaptation for Composite Web Services»; International Journal of Web Services Practices, Vol.4, No.1 (2009), pp.18-27.
- [9] D. Pellier, H. Fiorino. 2009. «Un modèle de composition automatique et distribuée de services Web par planification», RSTI-RIA-23/2009. Intelligence artificielle et Web intelligence, pages13 à 46.
- [10] N. Temglit, H. Aliane, M. Ahmed Nacer. 2008. «Un modèle de composition des services Web sémantiques»; Volume 11-Numéro spécial CARI2008 -Marc Kokou Assogba.
- [11] A. Yachir, Y. Amirat, A. Chibani, K. Tari. 2009. «Approche basée sur la qualité de service pour la composition automatique des services en robotique ubiquitaire»; Laboratoire Images, Signaux et Systèmes Intelligents—LISSI (EA3956)—Université Paris Est.
- [12] L. Zeng, B. Benatallah, M. Dumas, J. Kalagnanam, and Q. Z. Sheng. 2003. «Quality Driven Web Services Composition». In Proceedings of the 12th international conference on World Wide Web (WWW), Budapest, Hungary. ACM Press, May 2003.
- [13] L. Zeng, B. Benatallah, M. Dumas, J. Kalagnanam and H. Chang. 2004. «Qos-aware middleware for Web services composition». In IEEE Transaction on Software Engineering.
- [14] L. Li, M. Rong and G. Zhang. 2013. «A Web Service Composition Selection Approach based on Multi-Dimension QoS». The 8th International Conference on Computer Science & Education (ICCSE 2013) April 26-28, 2013. Colombo, Sri Lanka.
- [15] C. Jatoth and G.R. Gangadharan. 2015. «Fitness Metrics for QoS-Aware Web Service Composition Using Metaheuristics». Springer International Publishing Switzerland 2015, R. Neves-Silva et al. (eds.), Intelligent Decision Technologies, Smart Innovation, Systems and Technologies 39, DOI 10.1007/978-3-319-19857-6\_24.
- [16] Z. Zheng, Y. Zhang, and M. R. Lyu. 2014. «Investigating QoS of Real-World Web Services, IEEE Transactions on Services Computing». Vol. 7, No. 1, January-March 2014.
- [17] W. Khowfa, O. Silasai and C. Kaewpruksapimon. 2015. «QoS Based Service Selection in Cloud Environment: A Review». Int. J. Advance Soft Compu. Appl, Vol. 7, No. 3, November 2015 ISSN 2074-8523.

# A Novel Mapreduce Lift Association Rule Mining Algorithm (MRLAR) for Big Data

Nour E. Oweis

Faculty of Electrical Engineering and  
Computer Science, VSB-Technical  
University of Ostrava  
Ostrava, Czech Republic

Mohamed Mostafa Fouad

Arab Academy for Science,  
Technology, and Maritime Transport  
Cairo, Egypt

Sami R. Oweis

Alumni of Electrical and Computer  
Engineering,  
Oakland University  
Rochester, MI, USA

Suhail S. Owais

Department of Computer Science, FIT, Applied Science  
University  
Amman, Jordan

Vaclav Snasel

Faculty of Electrical Engineering and Computer Science,  
VSB-Technical University of Ostrava  
Ostrava, Czech Republic

**Abstract**—Big Data mining is an analytic process used to discover the hidden knowledge and patterns from a massive, complex, and multi-dimensional dataset. Single-processor's memory and CPU resources are very limited, which makes the algorithm performance ineffective. Recently, there has been renewed interest in using association rule mining (ARM) in Big Data to uncover relationships between what seems to be unrelated. However, the traditional discovery ARM techniques are unable to handle this huge amount of data. Therefore, there is a vital need to scalable and parallel strategies for ARM based on Big Data approaches. This paper develops a novel MapReduce framework for an association rule algorithm based on Lift interestingness measurement (MRLAR) which can handle massive datasets with a large number of nodes. The experimental result shows the efficiency of the proposed algorithm to measure the correlations between itemsets through integrating the uses of MapReduce and LIM instead of depending on confidence.

**Keywords**—Big Data; Data Mining; Association Rule; MapReduce; Lift Interesting Measurement

## I. INTRODUCTION

The recent advances in computers and communications have increased the number of relevant applications associated, such as a Radio Frequency Identification Devices (RFID), Wireless Sensor Networks (WSNs), Internet of Things (IoT), and other applications. Usually, these applications create a huge stream of non-stop data that is currently well-known as big data, denoted as “Big Data” [1]. Big Data is a massive set of data that is too complex to be managed by traditional applications. Nowadays, it includes huge, complex, and abundant structured, semi-structure, and unstructured data as well as hidden data that are generated and gathered from several fields and resources [2]. There are many challenges to manage such sets of Big Data include extracting, analyzing, visualizing, sharing, storage, transferring and searching [3]. These data (Big Data) are stored in powerful computers; include many of hidden patterns indicators that help in decision making. Data mining approaches facilitate decision making through determining and explain those patterns in a meaningful knowledge format [4].

Since, traditional data processing approaches and its applications could not be directly implanted when working with the big data management [5], it is necessary to apply new techniques, algorithms, and frameworks to manage, extract, and execute the big data mining development, and make these data mining techniques very helpful and more efficient.

Frequent pattern mining is one of the well-known data mining techniques that focused on discovering a number of interesting patterns from a large set of data items [6]. The association rule is a frequent pattern mining that is usually applied to find all the frequent co-occurrence relationships from a set of transactions [7]. Usually, the association rule strength is measured through two parameters (Support and Confidence). However these two parameters may not be sufficient to discover some interesting patterns, thus another measuring criteria is used which is the “Lift” [8].

On the other hand MapReduce is a parallel-based approach proposed for parallel processing of large datasets. This paper investigates the efficiency of the integration of both approaches (the Association Rule and the MapReduce). Therefore this paper proposed a parallel-based MapReduce approach for an association rule algorithm based on the Lift interestingness measurement.

This paper presents a MapReduce approach that has been used for defining the association rule importance based on the Lift interestingness measurement. This approach can be easily applied to many commodity machines to deal with big data. Finally, the work presented in this paper is in agreement with the published literature [1, 9, 10], and concludes that the traditional data processing tools and its applications are incapable of handling the current huge and complex data, such as, managing big data mining, and the newest industrial age of the IoT. Therefore, the parallel algorithm is the suitable solution for the big data mining techniques. This paper proposes a solution to solve one of the most critical problems of big data mining by emerging data mining, big data with parallelization and association rule to improve the usage of huge, complex amount of dataset.

Following the first section, this paper is organized as follows: Section two covers the background for association rule, and big data including the MapReduce paradigm. Section three covers the related work topics including the parallel association rule by utilizing several methods, measures, and techniques. Section four presents the proposed algorithm including the dataset, software use, and the novel parallel-based MapReduce approach for an association rule algorithm based on Lift interestingness measurement. Section five covers the experiments results and conclusion.

## II. BACKGROUND

This section presents the background information about the main topics used in this paper and the following subsection defines the association rule including different measured such as: support, confidence, and Lift interestingness measure.

### A. Association Rules

Data mining techniques contains a variety of applications and notable uses which are designed to work skillfully with a very huge amount of data. These applications and there notable uses cover wide domains of our life, including social networks, health care, financial, communications and many more.

Data mining approaches can be classified into two major models [11]: The descriptive data mining and the predictive data mining models. The descriptive models are unsupervised learning that describe the historical events, and the presumed or real relationship between elements that created them. These models uses a summarization analysis tools including multiple techniques such as, association rule for discovering and extracting relevant data. This model is commonly used in marketing analysis [11].

The predictive models are supervised learning that can accurately predict future outcomes based on existing data that carries out the analysis and extraction in more specifications and classifications. This model is commonly used in marketing predictions to forecast which new products may be popular in the future [12].

The association rules used within a dataset to discover non-trivial hidden patterns between items in a set could utilize either descriptive or predictive models [13]. In many cases, the algorithms generate large number of association rules, often in thousands or millions. It is almost impossible for users to visualize or validate such a large number of complex association rules, which limits the usefulness of data mining results. Therefore, it is important to identify that the components of an association rule are two sets of items: Left Hand Side (LHS) and Right Hand Side (RHS). The LHS is the antecedent (an item found in the transactions) and the RHS is the consequent (an item that is found in combination with the former). Moreover, there are two well-known measurements for an association rule, the Support of the rule and the Confidence of a rule [13].

#### 1) The Support Rule

The Support rule is considered a global measurement of interest for an itemset denoted by  $Supp(X)$ . The  $Supp(X)$  is calculated by counting the number of proportion transactions ( $P$ ) within the dataset as shown in (1) [14]:

$$Supp(X) = P(X \cup Y) \quad (1)$$

where:

$X$  is an itemset of interest.

$Y$  is an itemset with a defined condition of interest.

The itemset is called "frequent itemset" once its support output value is higher than a given minimum support.

#### 2) The Confidence Rule

The Confidence rule is considered a localization measure of correlation between  $X$  and  $Y$  denoted by  $Confidence(X \rightarrow Y)$ . The confidence rule is calculated as the ratio between the support of the union between  $X$  and  $Y$  subsets and the support of  $X$  as shown in (2) [15].

$$Confidence(X \rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)} \quad (2)$$

These two measures (Support and Confidence) may not be enough to extract some hidden patterns and to determine the correlation rule between LHS and RHS as mentioned in previous research work [16, 17, 18]. Therefore, an additional measurement is used; the Lift interestingness measure (LIM).

#### 3) The Lift Rule

Lift interestingness measure defines the number of transactions that contain the items used to find interesting patterns. The Lift measure is denoted by  $Lift(X \rightarrow Y)$  as shown in (3) [19].

$$Lift(X \rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X) \times Supp(Y)} \quad (3)$$

The Lift rule output defines the correlation between the LHS and RHS as follows:

- a)  $Lift > 1 \rightarrow$  Positive Correlation.
- b)  $Lift < 1 \rightarrow$  Negative Correlation.
- c)  $Lift = 1 \rightarrow$  Independent Correlation.

While most of the current algorithms proposed with regards to the association rule, such as Apriori [20], depends on either support, confidence or a combination of both rules, this paper utilized the "Lift measurement" instead of confidence to extract the association rules.

The proposed algorithm depends mainly on the support measurement, in contrast to other algorithms that depend on both support and confidence with Lift to extract association rules. Therefore, it simplifies the Lift measurement because it depends on support only.

### B. Big Data

Big Data is a complex, heterogeneous, massive, and hidden set of data that is hard to be managed, processed, analyses, and visualize by traditional applications. Gartner defined the term of big data as: "Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" [11, 21].

Although the Volume, Variety, and Velocity are the most common dimensions of the big data. There are other dimensions "Vs" that are recently defined such as: Veracity, Viability and value. While, the Volume describes the huge data capacity, the Velocity describes the speed of the data transmission and processing per interval time. The Variety defines the heteroge-

neity of data types. The Veracity focuses on the data quality (data cleaning from several noises). Finally the Viability and the Value dimensions: While the former defines the different data prediction possibilities, the later tends to gain valuable knowledge [22, 23]. Understanding these dimensions is important for designing big data mining techniques and platforms.

There are multiple big data analytic techniques used to extract, analyze, and visualize the complex and different data types. In the following section will introduce the MapReduce approaches as a main paradigm used in this paper.

### C. MapReduce:

MapReduce [24] is a data mining paradigm developed by Google that allows programmers to implement and processing large dataset with parallel and distributed algorithm on a cluster computing by using several programming languages such as: Matlab, C, C++, Java, Perl and more by using several MapReduce libraries [25, 26, 27].

MapReduce consists of two consequence processes, *Map* and *Reduce* and their functionally is defined as follows: The Map process is responsible for dividing, filtering, and sorting data tuples (key/value pairs) within using number of distributed clusters, the Reduce process summarizes the results into a few set of tuples [10, 28].

The advantages of MapReduce approach are many, for example: the big data classification approaches [29], the online Machine Learning for multicore and automatically failure handling [11, 30]. Parallelism also gives some possibilities partial recovery server failures: if the operating portion, which produces a pre-processing operation or convolution fails, its operation may be transferred to another working unit (assuming that the input data for the ongoing operation are available) and in others recent applications [31]. The most popular open source implementation of the MapReduce is the Apache Hadoop [32].

## III. RELATED WORK

The most well-known association rule algorithm is Apriori. Agrawal R. et al. [20] proposed the Apriori algorithm to extract relationships between data. This was done through applying a pruning technique to make the number of candidate itemsets much smaller and then find the frequent patterns to generate an association rule.

An improved version of Apriori algorithm was proposed by Aflori, and Craus [33]. It entitled the Frequent Pattern Growth (FP-Growth) algorithm capable for repeatedly reducing the search costs for short patterns. These patterns are linked to long frequent patterns to offer high selectivity mechanism.

Among the current research, there are several proposed parallel algorithms for association rule, such as the Parallel Apriori. This algorithm, proposed by Yang, X. Y., et al [34], is a parallel implementation of Apriori algorithm based on the MapReduce approach. The algorithm gives a solution to the exponential growth of data that encounter the traditional association rule mining techniques. This algorithm shows its benefits to deal with big data without consider the synchronization problem.

Jongwook Woo [35] proposed an Apriori-Map/Reduce algorithm and with both time and complexity, which theoretically shows that the algorithm provide much higher performance than the sequential algorithm as the map and reduce nodes get added. Also the paper shows that the itemsets produced by the algorithm can be adopted to compute and produce association rule for market analysis.

Chen, Y., et. al [36] The Parallel Randomized Algorithm for Approximate Association Rules Mining in MapReduce (PARMA) minimizes the data replication, the communication cost and the runtime improvement over parallel FP-Growth (FPF). The algorithm randomly separates the data into sets of samples. The machines works in parallel with their assigned set to produce deliverables and to be filtered and aggregated into a single output set.

Lin, X, et al. [37] proposed a parallel association rule algorithm called Niche-Aided Gene Expression Programming (NGEP). The advantage of the NGEP over both Apriori and FP-Growth is its efficiency to achieve more association rules with a higher accuracy rate.

Zhou, X., et al. [38] proposed an improved parallel association rules algorithm utilizing Hadoop as the MapReduce distributed programming framework. It has shown that the algorithm achieve well based on parallel performance and could be easily realized with the Hadoop platform.

Based on the previous literature, and briefly say that parallel association rule algorithms is one of the best choices for high performance big data mining techniques. Therefore this study proposes the application of the MapReduce approach for a paralyze association rule algorithm that is based on the Lift interestingness measurement.

## IV. THE PROPOSED ALGORITHM

As stated previously, the correlations between data items using the association rule measurements, is usually based on both confidence and support interests measurements. However the use of confidence is not effective to determine the association rules, since it does not describe the type of correlation(s) between the LHS and the RHS in the association rules.

The proposed algorithm that has been entitled "MapReduce-Based for Lift Association Rule (MRLAR)" is based on the Lift-Based Algorithm (LBA) [16] which is illustrated in Algorithm 1. Where the MRLAR improves the LBA algorithm through parallel executing. In which, the proposed algorithm works to determine the type of correlation between LHS and RHS in parallel association rules. The MRLAR algorithm was illustrated in Fig. 1.

The functionalities of the proposed algorithm MRLAR are discussed as follows:

- **Map function:** This step combines two steps; the data splitting step and the Mapping step [39]. The splitting step performed to distribute the data across each separated Map nodes. The map step consists of a map function that was established to find the association rules for some entities within a large sized database.



Algorithm 1: The Lift-Based Algorithm (LBA) [16].

1. Scans the database to find frequent items based-on minimum selections provided by the user
2. Choose correlation of association rule based-on Lift interestingness measure
  - a. If  $Lift > 1$ , positive correlation, Insert  $\alpha$ . (In this choice lift value  $> 1 + \alpha$ )
  - b. If  $Lift < 1$ , negative correlation, Insert  $\beta$ . (In this choice lift value  $< 1 - \beta$ )
  - c. If  $Lift = 1$ , dependent correlation, Insert  $\alpha$  and  $\beta$ . (In this choice lift value is between  $(1 - \beta$  and  $1 + \alpha)$ ).
  - d. Else, insert  $\alpha$  and  $\beta$ .
3. Second scan of the database
4. Get a frequent item for 1, 2 and 3 item by sequentially check for per value in item:
  - a. If  $\geq MinSupport$  added to frequent items
  - b. Else Ignore.
    - i. Generate candidate association rules from frequent items.
    - ii. Calculate the lift value for each candidate association rules to classify
5. Generate Association Rule for the choosing correlation of association rule.
6. If not found result (not found association rules), go to step 1 to edit MinSupport or edit the type of correlation).

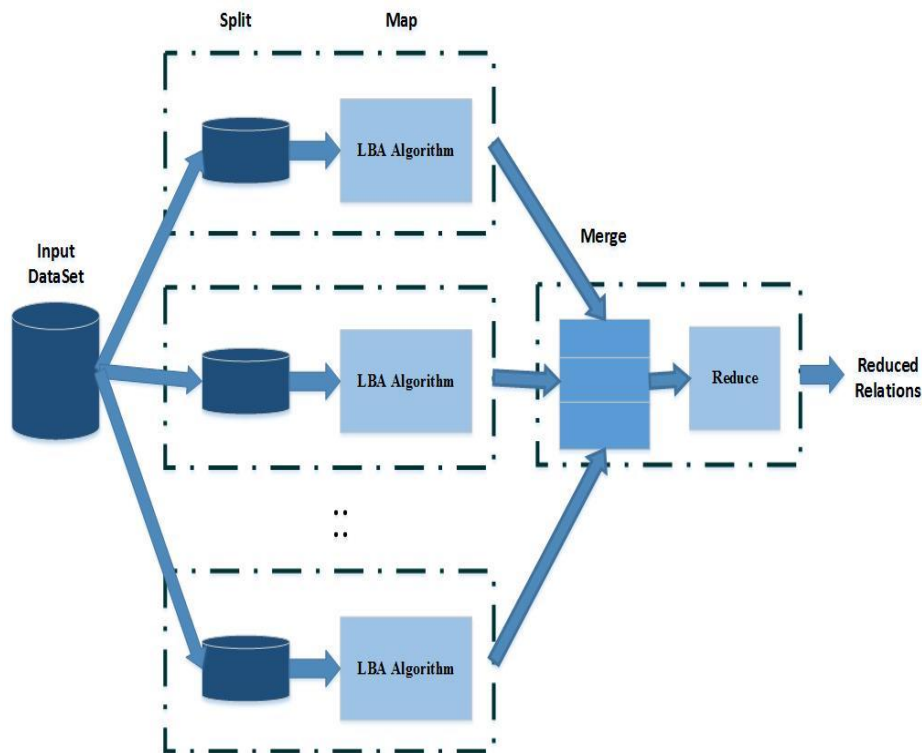


Fig. 1. Illustration of the MapReduce-Based for Lift Association Rule (MRLAR)

The Lift interestingness between some database key values is tended measurement within the MRLAR. Where selecting the key values from the user and based on the choices of the MRLAR can narrow down the search space in order to extract only the association rules of interest.

- **Reduce function:** The reduce process combines the outputs generated by each map node(s) to form the final collected association rules [37]. As it was mentioned previously the rules are weighted by the reduce process not the confident, but as an alternative in this parallel association rules where the authors involved the Lift weight computation to define the correlation between LHS and RHS.

The algorithm was designed and operated using MATLAB (R2015a) since it is well-known for its ability to support big data enhancement especially by using MapReduce approach [17]. As for experiments the algorithm used a dataset that was provided by the USA domestic airline flights between the period of 1987 and 2008.

The data comes originally from the Research and Innovative Technology Administration (RITA), the dataset is a large dataset: there are nearly 120 million records in total, and takes up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed with a collection of records consisting of 29 variables of flight information for several airline carries, including arrival and departure times with CSV files format.

These files have derivable variables removed, are packaged in yearly chunks and have been more heavily compressed than the originals. The full dataset can be downloaded from (<http://stat-computing.org/dataexpo/2009>)

In preparation for the collected data to remove irrelevant incomplete key values prior the startup of the MRLAR algorithm, two main processes were initiated; the data preprocessing and the data decoding.

- **Data Preprocessing:** Once the data was gathered from the RITA site and before it underwent the decoding process, preprocessing techniques was used to clean it up in order to make sure the data “datasotre” are free of empty cells, Not a Number (NaN) data, or any misrepresented data as strings which potentially could stop the execution of the program and/or result in wrong indication. The NaN and empty cells were assigned to zero.
- **Data Decoding:** Then the decoding process started, for Airport codes, which are identified as letters. Using MATLAB that able to find the unique list then assigning a unique digital code to every element of the list as shown in Table I column A. Another way was used to decode the strings by assigning a digit from 1-26 to each letter in the alphabets A-Z as shown in Table I column B.

TABLE I. ILLUSTRATION OF UNIQUE DIGITAL CODE AND STRING DECODING

A. Unique Digital Code		B. String Decoding by Digit	
Code	Decode	xStr	x_Code
ABE	100	A	1
ABI	101	B	2
ABQ	102	C	3
ABY	103	D	4
ACK	104	E	5
ACT	105	F	6
ACV	106	G	7
ACY	107	H	8
ADQ	108	I	9
AEX	109	J	10

Now since a digitalized file with only strings as the column headers, and used the “datastore” to read the file into memory. After that, scan the file for duplicated entries using the command “unique” when the data processing begin.

- **Initializing and key generation:** One of the most important reasons for designing a MapReduce algorithm based on Lift interestingness measure is to minimize the number of keys needed to be generated. This can be achieved by grouping the items by transactions. For example using “Day of Months” as a key from a collection of tabular dataset.

The Mapper function used to find the Lift method then pass this key-value pair to the reducer function. The reducer receives key-value pairs by key, and merges multiple cell arrays

with the same key into a single cell array. Subsequently can then store the result in a new datastore area. The whole MRLAR algorithm and its processing steps are illustrated in the Fig. 2.

## V. EXPERIMENTS AND RESULTS

Many of association rules use “Confidence” and “Support” measures for testing the occurrence of the itemset. Relying solely on both of them may be not sufficient. In this study algorithm, a test was setup as shown in Table II, for the association rules with their support, confidence, and lift measures with one attribute at a time (single-dimension association rule), that was “the delay in flights arrival of more than 60 minutes or more than 120 minutes”, also “the delay in flights departure of more than 30, 40, or 90 minutes”

Table II shows the reliability of the relationship between support and confidence at the single dimensionality level. For example, the first two tests ID numbers 1 and 2 in Table II showed that while the support drops from 0.045 to 0.013 the confidence also followed from 0.093 to 0.027 with a positive lift association’s value (22.3 and 75.95 respectively). Another point which can be noted is that the confidence increases while the number of itemsets increased.

The same results can be seen in ID numbers 3 and 4 with the same trend between confidence and support as well as the lift measure. However, when the data was challenged at the multidimensional level, the interpretation of the “Support and Confidence” can be misleading. For example, Items 3 and 4 in Table III showed inconsistency in the “support and confidence” trend. While the support has decreased in case 3 (Table III) from 0.0057 to 0.0044 in item 4 (Table III), the confidence has increased between these two cases (0.066 to 0.067). Hence, the use of “support and confidence” model only can survive under the single dimension association rule, but this is not applicable at the multidimensional level. Therefore, there is a need for a novel measure that is being able to be applicable at the multidimensional level, which is in this proposal the lift measure. It is important to state that relying on the lift adds benefits to the prediction process of the future consequence in future datasets with comparing to the current data.

In order to achieve more efficiency with high performance testing, parallel processing based on MapReduce was added in this study. The next experiment integrated multiple attributes with the same dataset. The experiments have been performed with the same two columns of interest (arrival delay and departure delay) but combined them with another attribute that was the month of the year (June and October). Table III shows the measurements for all support, confidence, and Lift measures (multi-dimensional association rule). The results showed that our approach has a high ability to run under several attributes, with the Lift interestingness measures successfully being able to determine the type of correlation between itemsets (positively, negatively, and independent) between LHS and RHS instead of using support and confidence.

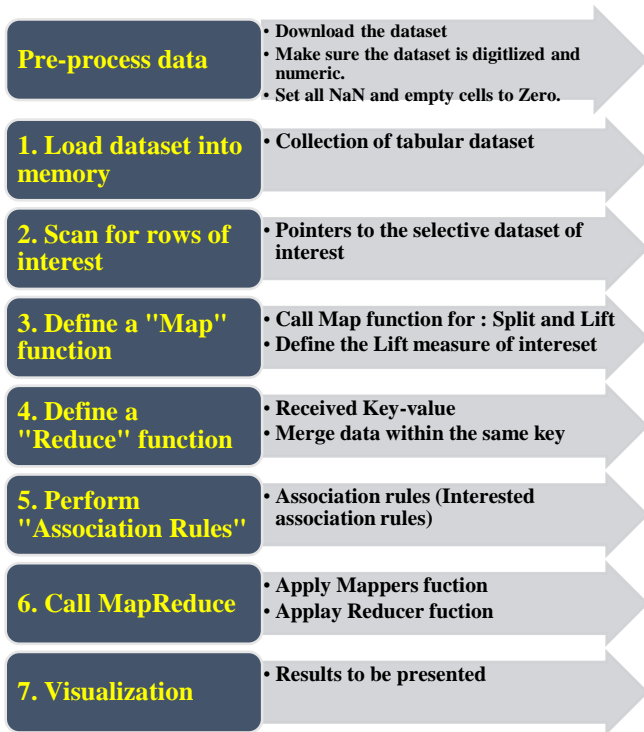


Fig. 2. Illustration of the MRLAR steps

TABLE II. ASSOCIATION RULES WITH THEIR SUPPORT, CONFIDENCE, AND LIFT MEASUREMENTS WITH ONE ATTRIBUTE TEST (ARRIVAL DELAY OR DEPARTURE DELAY)

a- Arrival Delay Cases (in minutes)					
ID	Rule	Itemsets	Support	Confidence	Lift
1	> 60	2946	0.045	0.093	22.3
2	> 120	865	0.013	0.027	75.96
b- Departure Delay Cases (in minutes)					
ID	Rule	Itemsets	Support	Confidence	Lift
3	> 30	5684	0.087	0.212	11.5586
4	> 40	4260	0.065	0.159	15.4223
5	> 90	1358	0.021	0.051	48.3792

By comparing both approaches single-dimension and multi-dimension association rules, Lift showed the ability to discriminate negative ( $< 1$ ) and positive ( $> 1$ ) relationships in the dataset when multi-dimension was applied. But in the other case with single-dimension, Lift always showed high positive values which did not presents a useful knowledge interpretation.

## VI. CONCLUSION

In this paper the authors developed a parallel association rule mining algorithm based on MapReduce paradigm by using Lift interestingness measured (MRLAR). The use of the MapReduce approach provided a powerful process over vast amounts of data utilizing parallel approach. Another measurable benefit that the MRLAR algorithm added its capability to directly extract association rule and type of correlation without the need to calculate confidence values, hence eliminating the need of additional calculations.

TABLE III. ASSOCIATION RULES WITH THEIR SUPPORT, CONFIDENCE, AND LIFT MEASURES WITH TWO ATTRIBUTES TEST (ARRIVAL DELAY, DEPARTURE DELAY AND MONTH OF THE YEAR)

a- Arrival Delay and October Cases					
ID	Rule	Itemsets	Support	Confidence	Lift
1	> 60 & Oct.	182	0.0028	0.0618	0.6678
2	> 120 & Oct.	54	0.0008	0.0624	0.6748
b- Departure Delay and October Cases					
ID	Rule	Itemsets	Support	Confidence	Lift
3	> 30 & Oct.	376	0.0057	0.0662	0.715
4	> 40 & Oct.	287	0.0044	0.0674	0.7282
5	> 90 & Oct.	87	0.0013	0.0641	0.6925
c- Arrival Delay and June Cases					
ID	Rule	Itemsets	Support	Confidence	Lift
1	> 60 & Jun.	266	0.004	0.0903	1.1883
2	> 120 & Jun.	99	0.0015	0.1145	1.5063
d- Departure Delay and June Cases					
ID	Rule	Itemsets	Support	Confidence	Lift
3	> 30 & Jun.	501	0.0076	0.0881	1.16
4	> 40 & Jun.	382	0.0058	0.0897	1.1802
5	> 90 & Jun.	141	0.0021	0.1038	1.3665

The experimental results presented in this paper show that MRLAR performs effectively the detection of associations between itemsets, through integrating the uses of MapReduce and Lift interestingness measured instead of using confidence to determine the correlation between LHS and RHS in association rule.

The proposed approach MRLAR showed a high ability to run under several attributes (multi-dimension association rules), also, Lift interestingness measure successfully were able to determine the type of correlation between itemsets (positively  $> 1$ , negatively  $< 1$ , and independent  $= 1$ ) between LHS and RHS instead of using confidence measure in a parallel association rules to help the user to make an important decisions making through determining and explain those patterns in a meaningful knowledge format.

## VII. FUTURE WORK

Although there are many benefits of MapReduce but it also have some limitations which cannot be passed over. Some of these limitations which are also specific to MRLAR such as an extracting the association rule with single dimension as shown in the results, also another limitation which is based on MapReduce approach itself that its operates only on data structures of type (key, value) pairs, so that all the input datasets must be adapted into such structure.

Another part of the future work is to develop a parallel data reduction techniques by using singular value decomposition (SVD) and semi discrete decomposition (SDD), or using any

other data reduction techniques to remove the unnecessary data with better time based on MapReduce approach. This algorithm lies in the ability to provide pre-processing techniques to reduce the dimensionality of the dataset which reduces the data capacity, for reducing costs. Hence, MapReduce dimensionality reduction algorithm by SVD may will handle another challenge of big data to avoid data dimensionality problems in parallel approach. Therefore, reduce the amount of time and memory required by data mining algorithms, easy visualization of data, and eliminate irrelevant features and noise reduction.

#### ACKNOWLEDGMENT

This Work is partially supported by Grant of SGS No. SP2016/97, VŠB - Technical University of Ostrava, Czech Republic.

#### REFERENCES

- [1] Oweis, N. E., Oweis, S. S., George, W., Suliman, M. G., & Snášel, V. "A Survey on Big Data, Mining: (Tools, Techniques, Applications and Notable Uses)". In Intelligent Data Analysis and Applications Springer International Publishing, pp. 109-119, 2015.
- [2] Fouad, M. M., Oweis, N. E., Gaber, T., Ahmed, M., & Snasel, V. "Data Mining and Fusion Techniques for WSNs as a Source of the Big Data". Procedia Computer Science, 65, ISSN 1877-0509, pp. 778-786, 2015.
- [3] Ding, Guoru, Qihui Wu, Jinlong Wang, and Yu-Dong Yao. "Big Spectrum Data: The New Resource for Cognitive Wireless Networking.", arXiv preprint arXiv: 1404.6508, 2014.
- [4] Larose, D. T. "Discovering knowledge in data: an introduction to data mining". John Wiley & Sons, 2014.
- [5] Saed Sayad, Data Mining Map, An Introduction to Data Mining, <http://www.saedsayad.com/>. (Last seen 1-Feb-2016).
- [6] Aggarwal, Charu C., and Jiawei Han, eds. "Frequent Pattern Mining". Springer, 2014.
- [7] Aggarwal, C. C. "Association Pattern Mining: Advanced Concepts". In Data Mining, Springer International Publishing, pp. 135-152, 2015.
- [8] McNicholas, P. D., Murphy, T. B., & O'Regan, M. Standardising the lift of an association rule. Computational Statistics & Data Analysis, 52(10), pp. 4712-4721, 2008.
- [9] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. "Data mining with big data. Knowledge and Data Engineering", IEEE Transactions on, 26(1), pp. 97-107, 2014.
- [10] Dhanshetti, A., & Rane, T. "A Survey on Efficient Big Data Clustering using MapReduce". Data Mining and Knowledge Engineering, 7(2), pp. 47-50, 2015.
- [11] Menon, S. P., & Hegde, N. P. "A survey of tools and applications in big data". In Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference, IEEE, pp. 1-7, 2015.
- [12] Zaki, M. J., & Meira Jr, W. "Data Mining and Analysis: Fundamental Concepts and Algorithms". Cambridge University Press, 2014.
- [13] Soysal, Ö. M. "Association rule mining with mostly associated sequential patterns". Expert Systems with applications, 42(5), pp. 2582-2592, 2015.
- [14] Chen, X. M., Wang, C. Y., & Cao, H. "Association Rules Mining Based on Minimal Generator of Frequent Closed Itemset". In Ecosystem Assessment and Fuzzy Systems Management, Springer International Publishing, pp. 275-282, 2014.
- [15] Makani, Z., Arora, S., & Kanikar, P. "A Parallel Approach to Combined Association Rule Mining. International Journal of Computer Applications", 62(15), pp. 7-13, 2013.
- [16] Hussein, N., Alashqur, A., & Sowon, B. "Using the interestingness measure lift to generate association rules. Journal of Advanced Computer Science & Technology", 4(1), pp. 156-162, 2015
- [17] Gancheva, V. "Market Basket Analysis of Beauty Products (Doctoral dissertation)", Thesis on Erasmus University Rotterdam, 2013.
- [18] McNicholas, P. D., Murphy, T. B., & O'Regan, M. Standardising the lift of an association rule. Computational Statistics & Data Analysis, 52(10), pp. 4712-4721, 2008.
- [19] Makani, Z., Arora, S., & Kanikar, P. "A Parallel Approach to Combined Association Rule Mining. International Journal of Computer Applications", 62(15), pp. 7-13, 2013.
- [20] Agrawal, R., & Srikant, R. "Fast algorithms for mining association rules". In Proc. 20th int. conf. very large data bases, VLDB Vol. 1215, pp. 487-499, 1994.
- [21] Srinivasa, S., & Bhatnagar, V. (Eds.), "Big Data Analytics", First International Conference, BDA 2012, New Delhi, India, December 24-26, 2012: Proceedings (Vol. 7678). Springer, 2012.
- [22] Kudyba, S. "Big Data, Mining, and Analytics: Components of Strategic Decision Making", CRC Press, 2014.
- [23] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data.", Information Sciences, vol. 275, pp. 314-347, 2014.
- [24] Dean, J., & Ghemawat, S. "MapReduce: simplified data processing on large clusters". Communications of the ACM, 51(1), pp. 107-113, 2008.
- [25] Koch, C. "Compilation and synthesis in big data analytics". In Big Data. Springer Berlin Heidelberg, pp. 6-6, 2013.
- [26] MathWorks Documentation, MapReduce <http://www.mathworks.com/help/matlab/ref/mapreducer.html?searchHighlight=mapreduce%20and%20matlab>, (Last seen 1-Feb-2016).
- [27] Fang, H., Zhang, Z., Wang, C. J., Daneshmand, M., Wang, C., & Wang, H. "A survey of big data research". IEEE network, 29(5), 6, 2015.
- [28] Triguero, I., Peralta, D., Bacardit, J., Garcia, S., & Herrera, F. "MRPR: A MapReduce solution for prototype reduction in big data classification neuro computing", 150, pp. 331-345, 2015.
- [29] Wang, B., Huang, S., Qiu, J., Liu, Y., & Wang, G. "Parallel online sequential extreme learning machine based on MapReduce Neuro computing", 149, pp. 224-232, 2015.
- [30] Giakoumakis, P., Chrysos, G., Dollas, A., & Papaefstathiou, I. "Acceleration of Data Streaming Classification using Reconfigurable Technology". In Applied Reconfigurable Computing, Springer International Publishing, pp. 357-364, 2015.
- [31] Bayramli, B. SVD Factorization for Tall-and-Fat Matrices on Map/Reduce Architectures. arXiv preprint arXiv:1310.4664, 2013
- [32] Sangavi, S., Vanmathi, A., Gayathri, R., Raju, R., Paul, P. V., & Dhavachelvan, P. An Enhanced DACHE Model for the MapReduce Environment. Procedia Computer Science, 50, 579-584, 2015
- [33] Wang, K., Tang, L., Han, J., & Liu, J. Top down fp-growth for association rule mining (pp. 334-340). Springer Berlin Heidelberg, 2002
- [34] Yang, X. Y., Liu, Z., & Fu, Y. "MapReduce as a programming model for association rules algorithm on Hadoop". In Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference, IEEE, pp. 99-102, 2010.
- [35] Woo, J. "Apriori-Map/Reduce Algorithm". In The 2012 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2012), Las Vegas, 2012.
- [36] Riondato, M., DeBrabant, J. A., Fonseca, R., & Upfal, E. "PARMA: a parallel randomized algorithm for approximate association rules mining in MapReduce". In Proceedings of the 21st ACM international conference on Information and knowledge management ACM. pp. 85-94, October 2012.
- [37] Lin, X. "MR-Apriori: Association Rules algorithm based on MapReduce". In Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference, IEEE, pp. 141-144, 2014.
- [38] Zhou, X., & Huang, Y.. "An improved parallel association rules algorithm based on MapReduce framework for big data". In Fuzzy Systems and Knowledge Discovery (FSKD), 2014 11th International Conference, IEEE, pp. 284-288, 2014.
- [39] Wang, B., Huang, S., Qiu, J., Liu, Y., & Wang, G. "Parallel online sequential extreme learning machine based on MapReduce Neuro computing", 149, pp. 224-232, 2015.

# Knowledge Management of Best Practices in a Collaborative Environment

Amal Al-Rasheed  
Information Systems  
Princess Nora University (PNU)  
Riyadh, Saudi Arabia

Jawad Berri  
Information Systems  
King Saud University (KSU)  
Riyadh, Saudi Arabia

**Abstract**—Identifying and sharing best practices in a domain means duplicating successes, which help people, learn from each other and reuse proven practices. Successful sharing of best practices can be accomplished by establishing a collaborative environment where users, experts and communities can interact and cooperate. A detailed review of previous research in best practice knowledge management shows that existing models have focused on developing methodologies to manage best practices but most of them did not propose solutions towards the development of full-fledged systems that make use of technologies to allow effective sharing and reuse of best practices. This paper presents a life cycle model to manage expertise for communities of practice. The proposed model is implemented in the education field as a knowledge management system that promotes and values user's contributions. We focus on the case of best teaching practice (BTP) as they develop instructor's abilities and improve overall instruction quality in higher education. For this purpose, we developed a computer environment including a knowledge management system and a web portal to assist instructors in higher education in the creation, sharing and application of BTPs.

**Keywords**—Best practice; knowledge management system; knowledge sharing; higher education; life cycle; portal

## I. INTRODUCTION

Knowledge is a key asset in the information age. Knowledge management can be seen as the set of activities that involve capturing all the best practices and knowledge that people acquire and storing them in a computer system in the hope that one day it will be useful [1]. The term “best practice” refers to “those practices that produce superior results; selected by a systematic process; and judged as exemplary, good, or successfully demonstrated” [2]. The purpose of managing BP is to learn from others' best or most successful performances and adjusting their approach to meet specific needs [3]. In fact, deciding what is “best” is not easy. Best practices can vary over time, as new evidence and new possibilities arise, and from place to place, depending on available resources and infrastructure [4].

People who share common interests connect with each other by participating in a community. Thus, successful sharing of BP can be accomplished by establishing a well-motivated community [5, 6]. A common definition of community emerged as a group of people with varied characteristics who share common perspectives, and engage in joint action in geographical location [7]. The concept of community has changed since the advent of the Internet. Now, individuals who

share common interests can meet virtually in an online community without geographical restriction [8]. The great majority of virtual communities depend on Web-based environments to connect, collaborate, communicate and contribute to community activities [9, 10]. A community's specific purpose and goals inform the appropriate activities and technologies that should support it.

In this paper, we propose a BP development life cycle that specifies the required steps for engineering BP and integrates the Web technologies that are essential for their implementation. Application of the proposed model is illustrated by reference to the education field—specifically, the use of best teaching practices (BTP) to develop instructor abilities and improve overall learning in higher education.

The rest of this paper is structured as follows. The next section explores the concept of BP, and we then introduce our proposal for a BP life cycle model. The model is illustrated by applying it to the education field, and the paper ends with a brief summary and conclusion.

## II. BEST PRACTICE IN KNOWLEDGE MANAGEMENT

### A. What is a “best practice”?

A best practice is generally defined as “an improvement in a particular process, approach, technique, or subject matter knowledge that is good enough to replace an existing practice and general enough to merit being disseminated widely throughout an organization” [11]. The term is used in many areas that include health, education, government administration and project management [12]. Some people favor the term “good practice” on the grounds that while there are many good practices, it is difficult to achieve consensus on a single “best practice”. An acceptable alternative definition of BP, then, is “one that has been proven to work well and give good results, and is therefore recommended as a model” [13, 14].

It is important to note that most BP-related knowledge is tacit; this can be difficult to document because it is held in people's heads and may therefore be less easily expressed [15]. For this reason, BP-related programs necessarily include two key components: explicit knowledge and tacit knowledge. While sharing of explicit knowledge connects people with information (e.g., in a BP database), tacit knowledge sharing connects people with other people (e.g., through communities of practice). These methods complement each other; while a database enables users to search for BP, the best way of

sharing BP is by joining communities and interacting with individuals who have previously used that BP [16].

### B. The benefits of sharing BPs

Replicating a success in one part of the organization across all others is one of the best ways to improve performance [17, 18]. Sharing internal BPs is particularly helpful when an organization comprises multiple units, enabling individuals doing similar work to gain from each other [19]. In organizations, identifying BPs enhances learning and reuse of proven practices among staff members. Additionally, successful sharing of BPs enables an organization to i) improve the quality of services provided; ii) avoid duplication of effort or “reinventing the wheel”; iii) reduce the need to redo work; and iv) save money through increased efficiency and productivity [20].

## III. MANAGING BEST PRACTICES

The integration of BPs within an organization is not a straightforward task and is generally implemented as a life cycle. A life cycle introduces specific steps for understanding how organizational knowledge is processed during its useful lifespan

### A. Previous work

A review of the relevant literature identified a number of models of BP. The “pull” model as a mechanism for BP transfer was proposed by Pickard and Golden (1995). Its basic idea was to establish a technology transfer agent and then to assign a technology transfer mechanism to users in that industry sector by means of interviews and workshops. The main characteristic of the “pull” model is a high level of face-to-face communication between technology suppliers and industry executives [21].

Ashton (1998) was among the first to address the need for a “coherent and practical framework for BP,” which he sought to create by identifying a set of process phases [22]. He found that the organizations under consideration already had a defined BP management process. Ashton noted the importance of technology in improving the corporate BP process, and of measuring it against credible frameworks.

Jarrar & Zairi (2000) made a significant contribution by constructing a framework based on success factors derived from the experiences of various organizations [2]. As well as the framework, they emphasized that advances in IT would be the main factor to impact benchmarking in the future.

In another important study, Signal (2006) constructed a BP transfer process consisting of five main steps [18]. The first of these was to determine critical processes and factors relating to quality, safety and productivity. The next step was to prepare drafts for stakeholders and process experts, detailing BP for the processes in question. The drafts would then be reviewed

before finally accepting and publishing the reviewed drafts as BP benchmarks.

Another framework in [23] suggested six major steps for identifying and sharing BPs. The general aim of this approach was to define the significant features of a BP, identifying the relevant experts, concluding overall guidelines, diffusing basic knowledge and encouraging subject matter experts to apply and adjust the practices in alignment with the new context.

### B. The BP Cycle Model

While each of the above frameworks clearly introduced valuable new elements to the BP process, these earlier approaches have the following limitations:

- Many frameworks applied to traditional communities in which members generally meet face-to-face; current technologies extend this concept by overriding geographical and social boundaries, removing language barriers and bringing together experts from around the world [24].
- Earlier approaches focused on acquiring explicit BP through phases such as identify, find, search and discover. However, most BP knowledge is tacit and requires specific mechanisms to capture it. In practice, tacit knowledge is hard to acquire because it is difficult to communicate to others. Consequently, the creation of new knowledge entails the establishment of a virtual community and the fostering of social interactions among its members [25].
- Previous models focused on developing methodologies to manage BPs, but few have proposed solutions for the development of fully-fledged systems that exploit technologies for effective sharing and reuse of BPs.

Several important phases from earlier frameworks and life cycles have been adopted and incorporated in the construction of our simple, practical and comprehensive BP life cycle model for virtual communities. The aim of this model is to transform the individual knowledge residing in people’s heads into public knowledge that is available to all practitioners. As illustrated in Fig. 1, the BP Cycle (BPC) model comprises six cyclical phases.

In some knowledge management cycles, the *identify* and *create* phases are grouped together; some authors have argued for this approach on grounds of the clear overlap between the two [26]. In our opinion, all practices have to pass through these as two distinct phases. When a BP request is made, the searcher must first identify the appropriate knowledge if it exists in-house; if not, the BP must be created or acquired. Even if the required knowledge is found, in most cases it will need to be tailored for each specific situation by auditing existing BPs [27].



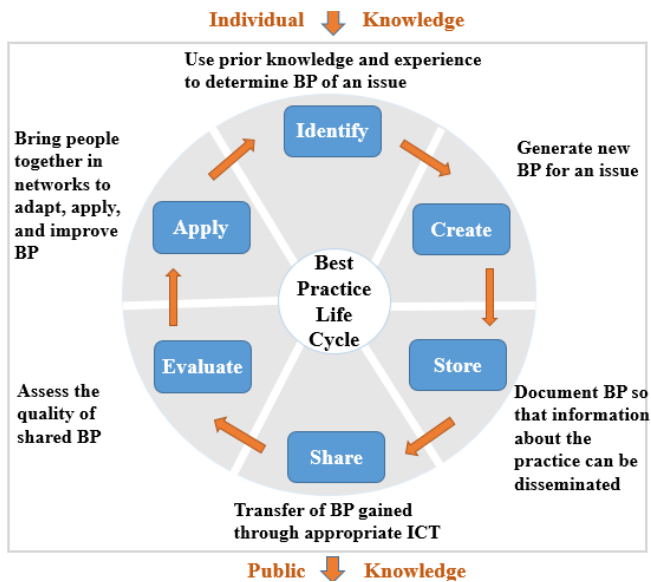


Fig. 1. The BP Cycle (BPC) Model

The key steps of the BPC model are as follows.

#### 1) Identify

This phase is about searching and finding best practices wherever they are. However, finding best practices is not easy. It involve using prior knowledge and experience to determine best ideas/practices of an issue. Identifying BPs focuses on explicit knowledge that can be found, codified and formalized, in sources such as literature reviews, databases, knowledge repositories and training courses [2]. Actually, this phase starts by looking at what areas or issues need attention, considering where one can really add value, and determining who will benefit most from better knowledge and understanding of good practices [2, 23]. Along with effectively finding best practices, the identify phase implicates analyzing and assessing knowledge based on specific evaluation criteria [26].

#### 2) Create

This phase involves generating new good practices for an issue if none found through searching during the identify phase. New best practices may also need to be created if existing ones only partially satisfy knowledge needs [26]. Much of best practice knowledge is tacit and not always easy to document [16]. Therefore, this phase implicate using some tools and technologies for elicitation tacit knowledge. There are several ways to create good practices. One is to examine individuals and groups that provide excellent results and therefore expected to be using good practices. Having discovered these, one will then need to recognize what parts of their whole method or approach represent good practice. However, other approaches exist too such as communities of practice and interviews. In fact, knowledge workers learn from their experience. They may create new knowledge as a response to the “failures or successes” of previous phases. This way, feedback is used to create new knowledge that has been contextualized [1].

#### 3) Store

In this phase, the codified best practices are stored physically, which entails using technology to build knowledge

repository. Some examples of technology include intranets; databases; knowledge portals; archives of knowledge; and information systems [28]. The repository cannot be a random collection of best practices regardless of their individual and collective value. Best practices must be stored in a structured way that allows them to be manipulated, retrieved, and shared efficiently. Common related activities include tagging, annotating, archiving, classifying, and improving search and retrieval needs [26]. Moreover, to store best practices effectively, their descriptions are commonly kept in a database in a standard format. A typical template might contain the following information about best practice: title, profile, context, resources, description, lessons learned, links to resources, and tools and techniques [23].

#### 4) Share

This phase represents the core of ‘best practice life cycle’ since the primary goal of the process is to make the contributed best practices available for individuals to take advantage of them. ‘Share’ entails internal and external collaboration and communication between individuals, along with extensive use of technology for the dissemination of knowledge. An explicit and flexible network of expertise (as in a community of practice) enhances collaboration and can strongly assist the sharing of BPs [26]. It is also important to choose the optimum mix of technologies and dissemination channels [11]. Some of the most common technologies used to share knowledge assets are content management systems, communication and collaboration technologies, blogs, social media and websites [11].

#### 5) Evaluate

A practice is judged to be good if there is an obvious link between what is practiced and the desired outcome [23]. In most cases, assessing what we believe a good practice is essential. Meantime, deciding on the value of different ideas in relation to the required intention and the outcome in question. The evaluate phase is context-based and dependent on the specific situation and individual or group, who must validate BPs based on their experience [2]. Such assessments of validity are based on two questions: Is a practice perceived as BP really best, and does the transferred BP perform as required? [2]. A common approach is to ask a panel of peer reviewers to evaluate a potentially good practice [29]. Moreover, some organizations tackle the validity issue through iterative process of workshops, guidelines featuring assessment, feedback and measurable improvements. These ensure validity through assessment and feedback, determine if the best practices have produced measurable improvement, and consider whether it is recognized by internal and external sources. There is no doubt that more extensive use of information technology and social networks offers less expensive methods for assessing information and knowledge in general. Such methods depend on community interaction and feedback by motivating users to add their ratings and comments.

#### 6) Apply

The importance of ‘apply’ lies in the fact that it is the motive for the evaluation and creation of more knowledge. In this phase, people is brought together in networks to adapt, apply, and improve best practices. Once shared, BPs can be activated and their value extracted and applied to solve



problems, make decisions, improve efficiency or promote innovative thinking [26]. This phase is also key to internalizing tacit forms of knowledge and for that reason is sometimes called “learning-by-doing” [11]. Some of the more common activities in the apply phase include developing communities of practice, locating experts and running coaching workshops and tutorials. The technologies employed in these activities may include knowledge repositories, search engines and communication and collaboration technologies [28]. It is important to note that all of the KM efforts have been in vain unless this phase is performed successfully [26]; application of knowledge is the key success of any knowledge management system (KMS).

Surrounding the process, and helping it, what we call the enablers technology, culture, leadership, and measures. We have to address the aspects of the infrastructure and the environment of the organization in order to enable the transfer process to have a change in work. One reason that the internal transfer is so difficult is that these enablers have been poorly understood and were rarely addressed in earlier attempts [17].

#### IV. MANAGING BEST PRACTICES

In conducting this research, we have focused on best teaching practices (BTP) that develop instructor abilities and improve overall learning in higher education. BTP represents instructors’ expertise as accumulated while teaching and can be defined as a “method that clearly adds value in teaching a course by saving time or clarifying a concept.” BTP is acquired through experience; it has been applied many times and has been proven to work well, gives good results and can therefore be recommended as a model. Identifying, sharing and reusing BTPs means duplicating successes to help instructors to learn from each other, resulting in innovative ideas for improving the effectiveness of teaching and learning. BTP is a key influence on student learning, and as such, it is a desired outcome and an essential objective of higher educational institutions. University instructors often struggle to adhere to the principles of BTP in attempting to provide the best learning experience for their students [30, 31].

In the e-learning field, most efforts have focused on the expression and transmission of course content to learners, with very little attention given to transmitting instructors’ expertise or enabling instructors to communicate their methods of teaching and delivering knowledge. It is essential to create and provide an appropriate environment and technical conditions for instructors to create, transfer, share and apply knowledge effectively [8]. Clearly, this should include exploiting, improving and managing the existing BP knowledge of university instructors to enhance the overall development of teaching and scientific research levels in higher education.

#### V. APPLICATION OF THE PROPOSED BPC MODEL

This section illustrates the use of the proposed cycle in the context of BTP.

##### A. Identify and Create BTP

In the proposed model, the process of identification and creation of BP knowledge entails seven steps (Fig. 2).

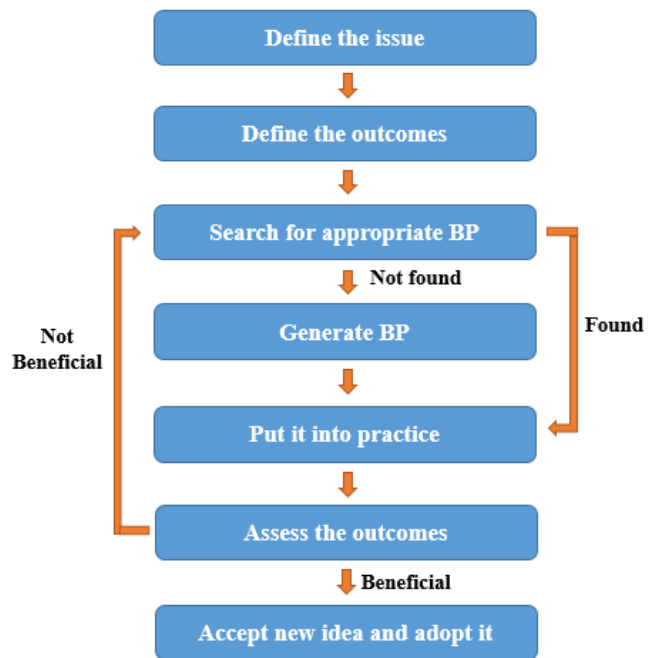


Fig. 2. Steps for identification and creation of BP knowledge

##### 1) Define the issue

The need for good/best practice stems from the existence of a difficulty facing an instructor in teaching a specific topic. Defining clear objectives is key to the success of identifying and creating BP. Defining the issue precisely eases the process of finding resources that match the intended objective.

##### 2) Define the outcomes

The instructor must define the outcomes they intend to achieve in applying a BP. The selected BP should aim at accomplishing the predefined outcomes. However, defining outcomes guide the instructor in developing appropriate learning experiences for students, and play a major role in assessment.

##### 3) Search for appropriate BP

Once the instructor has defined what they are looking for, they must establish what is available. When faced with a number of relevant BP choices, the next step is to minimize the search results by eliminating those that are unsuitable for the students in question or do not address the intended outcomes. Instructors can find BPs using:

- The Internet. The Internet is undoubtedly the largest single source of best practice information available. Anyone with good search skills can find nearly anything.
- Networking. Contacting knowledgeable people in the same field to find out what they know is useful in inspiring us and enriching our information.
- Libraries. Libraries are a great source and helpful in finding what we are looking for. Now, we can find many journals and individual journal articles on the Internet as well.

- Communities. Clearly, members of the same community may know of successful practices or initiatives similar to the one that we need.

#### 4) Generate BP

If no relevant BP is found by searching or if those found only partially satisfy knowledge needs, the instructor must generate a new BP. As much BP knowledge is tacit and may be difficult to generate and document, the instructor may draw on prior experience or consult with other academics in the same field to benefit from their successful practices.

#### 5) Put it into practice

Knowledge and new ideas are of no value unless we put them into practice. On finding or generating BPs, the instructor should apply them in teaching to test their applicability.

#### 6) Assess the outcomes

Outcome assessment aims to measure the success and applicability of new ideas. We may consider new ideas to be good or best practices if they prove their success following implementation. The instructor should compare the results against intended outcomes and then use the resulting information to improve the quality of the new ideas. If the issue remains unresolved, return to the starting point.

#### 7) Accept new ideas and adopt them

Once a new idea has met expectations, it may be regarded as a good or best practice and applied in teaching. However, even a practice that is considered good or best may be further developed, and one should always remain open to possible improvements.

### B. Store BTP

Collected BPs can usefully be stored in an accessible knowledge base for future reuse. Today, we can access extensive knowledge of great value. The challenge is not the creation of new knowledge; the problem is that existing knowledge is often badly organized and almost inaccessible.

The issue, then, is how BP information can be organized and classified in a knowledge base to enable users to more easily find and share what they need. Organizing and classifying BP knowledge involves its formal specification, which defines the attributes of knowledge and how it will be represented. Defining knowledge attributes is the first step toward designing a knowledge management system.

Knowledge attributes describe the artifacts exchanged and shared by users to refer to information obtained by the knowledge creator. In practice, this means that experts provide information manually, and because these attributes cover a wide range of information, it is difficult to define precisely its scope. The information gathered about best teaching practices falls into two main categories: knowledge object and knowledge creator or consumer.

#### 1) Knowledge object (KO)

The many definitions of knowledge to be learned include terms such as “knowledge objects” [32], “educational software components” [33], “sharable content object” [34], “courseware unit” [35], “educational objects” [36] and “learning object” [37]. In considering best teaching practices, the broadly used

concept of KO [32] seems most convenient. According to Merrill (1998), a KO is “a precise way to describe the subject matter content or knowledge to be taught” and a “way to organize a data base (knowledge base) of content resources (text, audio, video, and graphics)”. The components of KOs should not be confined to a particular subject area; the same components can be used to represent knowledge in different areas (e.g., science, mathematics, humanities, and technical skills).

In our research, we addressed learning object metadata, as these have much in common with teaching knowledge. In general, two popular approaches are used to describe metadata; the first collects metadata as records (implemented by the LOM standard) while the second considers metadata items individually (adopted by Dublin Core) [38]. For the purposes of a very general model and to ensure compatibility with existing platforms, we have adopted the second approach. However, as Dublin Core standards are used to describe the “learning object,” we have added new attributes to fit with KO description and the BP process. Relevant attributes of KOs include keywords, material type, course name, course level, discipline, sub-discipline, attachment and usage. Table 1 illustrates the main attributes of the KO and provides a detailed description of those attributes.

TABLE I. DESCRIPTION OF ATTRIBUTES OF KO

Attribute	Description
Title *	Name of KO as given by its creator
Description *	Textual description of KO content
Keywords	Tags describing the KO topic
BP type	Classification of KO by its specific use or application (e.g., teaching experience, lessons learned, user guide)
BP (Subject) *	Detailed description of the topic
Creator *	Person responsible for making the content
Discipline	Branch of knowledge to which the KO belongs (e.g., computer science, engineering, mathematics)
Sub-discipline	Field of specialized study within a discipline
Course name	Name of course that includes the KO topic
Level of course	Study level of course that includes the KO topic
Coverage *	Time, culture, or region to which the KO applies
Language *	Primary language used within the KO
Attachment	Related file/s (e.g., MS word, PowerPoint, PDF)
Media format *	Technical type of the learning object (image, audio, video)
Source *	String used to access the KO (e.g., Universal Resource Locator)
Publisher *	Entity responsible for making the content available
Contributor *	Person responsible for contributing to the content
Date *	Date of creation or availability of the content
Identifier *	Referring to the KO
Usage	Flag indicating the KO's use in teaching
Rights *	Terms of use of the KO

\* Source: Dublin Core Standards

2) Knowledge creator/consumer

The knowledge creator is the person responsible for contributing new knowledge based on their experience in a particular field; the knowledge consumer uses the new knowledge to supplement their existing knowledge. The instructor’s information should be taken into account to facilitate knowledge sharing and dissemination. Table 2 details the main attributes of the knowledge creator/consumer.

TABLE II. DESCRIPTION OF ATTRIBUTES OF KNOWLEDGE CREATOR/ CONSUMER

Category	Attribute	Description
General	First Name	Instructor’s first name
	Last Name	Instructor’s last name
	ID	Unique label that identifies the instructor
	Photo	Personal photo
	E-mail	Instructor’s e-mail
	Website	Instructor’s website address
	Bio	Detailed description of instructor’s life, including education, work, relationships, and experience
Profession	Title	e.g., PhD, professor, engineer
	Educational level	Instructor’s degree
	Institution	Institution that the instructor belongs to
	Department	Department that the instructor belongs to
	Discipline	Branch of knowledge studied by the instructor in HE education
	Sub-discipline	Field of specialized study within
	Area of expertise	List of the instructor’s teaching areas
	Courses	List of courses that the instructor teaches

Fig. 3 provides a detailed example of information about BTP stored in the knowledge base.

C. Share BTP

Instructors should be able to express their teaching innovations. The main challenge is to persuade instructors to share their expertise and to interact with their peers. As mentioned above, constructing the knowledge base is the first step toward designing a knowledge management system to support the knowledge base and to facilitate sharing and dissemination of existing knowledge.

To that end, we have developed a KMS that supports and stimulates instructors, educators and communities to express and share their expertise. The proposed system is based on two key principles: i) free-user contribution, which means that any proposed contribution by any instructor is welcome and there is no control or moderation of contributions by any instructor; ii) peer scoring of knowledge, which means that any contribution will be scored by other users, achieving a high score if it is frequently reused, liked and positively commented. The purpose of these two principles is to encourage voluntary contributions and to ensure fair evaluation and feedback from specialists. The proposed system promotes parity of interaction, where users work in a collaborative environment to promote co-construction and sharing of knowledge resources, with direct advantages for participants.

The KMS was implemented as a knowledge portal using Drupal (drupal.org), which is a free community-supported content management system for creating, organizing, presenting and managing a website [39]. Drupal offers more flexibility for the programmer than other existing content management systems to develop user specific applications. The portal runs on a digital platform that supports Apache, PHP and MySQL to store content and settings.

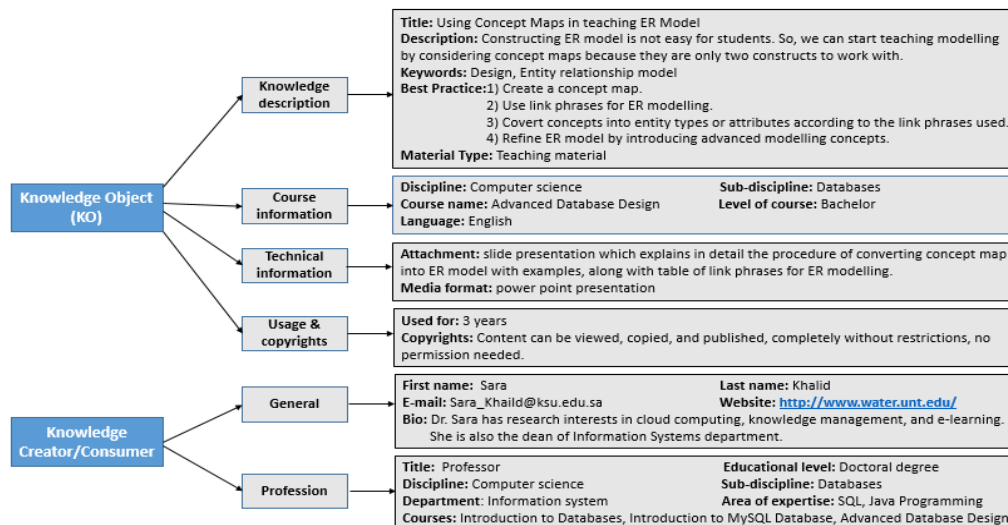


Fig. 3. Example of stored information about BTP

The knowledge portal maintains a dynamic client-side graphical user interface that handles all user requests and collaborative activities. It facilitates knowledge acquisition, discovery and sharing by enabling instructors to publish knowledge, work collaboratively, share ideas and store

knowledge resources in a searchable knowledge base. The knowledge portal also authenticates users and manages the rights and authorizations for each user within a more broadly defined university community. Fig. 4 illustrates the knowledge portal homepage.



Fig. 4. Knowledge portal homepage

In the portal, many techniques were adopted to encourage the community to become more involved and self-supporting. Discussion forums were set up to foster member engagement and sharing of expertise and information. Members from the same discipline share a dedicated forum; these currently include computer science, mathematics and engineering. The portal also concentrates on capturing and sharing knowhow expertise through blogging, where instructors with the same interest form groups to contribute and post their experiences. Unlike forums, blogs are established on the basis of interests rather than disciplines. A list of communities is illustrated in Fig. 5.

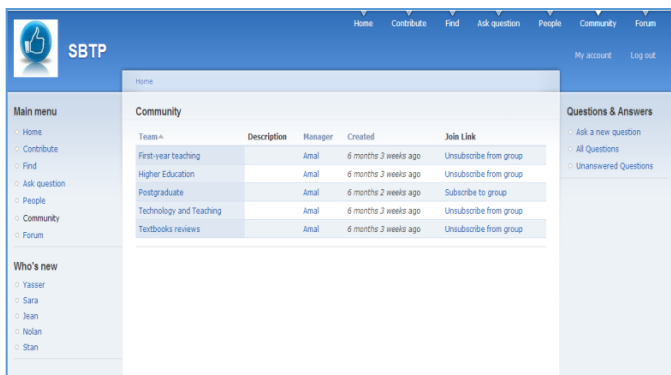


Fig. 5. List of communities

#### D. Evaluate BTP

The increasing volume of knowledge makes it more difficult to access target knowledge. Most searches of knowledge bases return a large number of KOs without any indication of quality or value, making it difficult for users to identify results that meet their needs. This creates an urgent need to find new ways of evaluating and presenting KOs ranked by quality and value. Existing evaluation tools are costly and time-consuming because they are qualitative and rely on expert review [40, 41], and there is a need for new tools that automatically evaluate and provide an assessment of knowledge resources, based on a number of specified parameters. Evaluation tools involve collaboration among multiple users, in which participants converge on accurate evaluations through a combination of user ratings, comments and views. For that reason, the proposed system automatically evaluates knowledge, assesses its quality, recommends the qualified experience in terms of various measures and assigns a

quantitative score for overall rating of KOs, allowing search results to be re-ranked according to the quality score.

The system adopts an extrinsic motivator, based on a reward system for scoring BTPs. The scoring function combines three weighted component measures: i) author score, measuring the relevance of an author in terms of their contribution to the system; ii) feedback, representing the degree of satisfaction of other users who have shared and reused the BTP; and iii) comments, measuring the feedback of users who did not use the BTP. The scoring function is a reward system that measures both contributor importance and the reputation of the BTP. When community members interact with a specific BTP as users or simply by adding comments, they automatically affect its rating. Earning reputation points by contributing knowledge can encourage an individual to become more involved when they recognize these benefits.

#### E. Apply BTP

The key to the success of any KMS is the reuse and application of stored knowledge. When that knowledge is used, it is refined and developed, and additional knowledge is created; the knowledge base is enriched; and the cycle repeats itself. Belief in the effectiveness of BPs is a prerequisite for replication; without this, the system will not succeed, even if everything else is in order. It follows that members must be persuaded that the adoption, use and implementation of BPs is feasible. This can be achieved by educating the community about the importance of BPs and by stimulating their commitment to use them. To do so, our system provides person-to-person linkage to support the application and use of tacit knowledge. Additionally, links connect members to experts who have identified and used the BP for further information. This technique helps to improve the transfer of tacit knowledge and social community practices among members.

## VI. CONCLUSION

Knowledge management (KM) is concerned with finding ways to make tacit knowledge explicit through documenting best practice to enhance knowledge sharing through human to human collaboration. Consequently, the best practice system is recognized as one of the main KM solutions in many KM literature or research. The first contribution of this paper was the construction of a practical and comprehensive best practice life cycle model for communities of practice which comprises six phases: identify create, store, share, evaluate, and apply. This model is based on studying earlier best practice processes and life cycles in the literature. The second contribution of this research is the development of a system that allows communities of practice who share common interests to build a collaborative environment to create, disseminate and apply best teaching practices in teaching. Creating an online community where university instructors would freely share their innovative teaching experiences can leverage noticeably the quality of instruction for the benefit of students and instructors. The implemented system has been used by a community of instructors at our university to interact, collaborate and share BPs within their fields of interest. The results of this experience are very promising and encourage further development of this research.

Future work will investigate the possibility of coupling our system with the learning management system (LMS) currently used by the university's instructors.

This will allow instructors to input BPTs automatically from the LMS in an integrated learning environment that fosters more intuitive exchange and collaboration.

#### REFERENCES

- [1] R. Magnier-Watanabe, C. Benton, a. countries and D. Senoo, "A study of knowledge management enablers across countries," *Knowledge Management Research & Practice*, vol. 9, p. 17–28, 2011.
- [2] Y. Jarrar and M. Zairi, "Best practice transfer for future competitiveness: a study of best practices," *TOTAL QUALITY MANAGEMENT*, pp. VOL. 11, NO. 4/5&6, p.734- 740, 2000(a).
- [3] R. Cox and C. Cunial, "Best Practice Within Australian Food Service, a Case Study: Customer satisfaction with red meat products," *Australian Agribusiness Review*, p. 14 (paper 13), 2006.
- [4] Y. F. Jarrar and M. Zairi, "Internal Transfer of Best Practice for Performance Excellence: A Global Survey," *Benchmarking: An International Journal*, pp. Vol 7, No 4, pp. 239-246, 2000(b).
- [5] F. Henri and B. Pudelko, "Understanding and analysing activity and learning in virtual communities," *Journal of Computer Assisted Learning* (2003), pp. vol. 19, pp. 474-487, 2003.
- [6] R. Ramasami, *Knowledge Management*, Lulu.com, 2011.
- [7] K. M. MacQueen, E. McLellan, D. S. Metzger, S. Kegeles, R. P. Strauss, R. Scotti, L. Blanchard and R. T. Trotter, "What is community? An evidence-based definition for participatory public health," *American Journal of Public Health*, pp. vol. 91, no. 12, pp. 1929–1938, 20, 2001.
- [8] J. Hautala, "International academic knowledge creation and ba. A case study from Finland," *Knowledge Management Research & Practice*, vol. 9, p. 4–16, 2011.
- [9] D. Cambridge, S. Kaplan and V. Suter, "Community of Practice Design Guide," *Communities*, pp. 1–8, 2005.
- [10] S. Dani I, J. A. Harding, K. Case, R. I. M. Young, S. Cochrane, J. Gao and D. Baxter, "A Methodology for Best Practice Knowledge Management," *Proceedings of the Institution of Mechanical Engineers, part B: Journal of Engineering Manufacture*, , pp. 220 (10), pp. 1717-1728, 2006.
- [11] K. Dalkir, *Knowledge Management in Theory and Practice*, Oxford: Elsevier, 2013.
- [12] WHO, "Guide for Documenting and Sharing Best Practices in Health Programmes," *World Health Organization, Regional Office for Africa, Brazzaville*, 2008.
- [13] FAO, "Good practices at FAO: Experience capitalization for continuous learning," *Food and Agriculture Organization of the United Nations (FAO)*, 2013.
- [14] S. G. Rabinowicz and V. Chinapah, "Good Practices in Pursuit of Sustainable Rural Transformation," *Journal of Education and Research*, pp. Vol. 4, No. 2, pp. 3-19, 2014.
- [15] W. Reddy and S. McCarthy, "Sharing best practice," *International Journal of Health Care Quality Assurance*, pp. vol. 19, no.7, pp. 594-598, 2006.
- [16] G. Servin, "ABC of Knowledge Management," *NHS National Library for Health: Knowledge Management Specialist Library*, 2005.
- [17] C. O'dell and J. Grayson, "Identifying and Transferring Internal Best Practice," *International Handbooks on Information Systems* , pp. Volume 1, pp 601-622, 2004.
- [18] P. Signal, "Best Practice Makes Perfect (2006)," *Dairy Industries International*, pp. Vol. 71, No. 2, pp. 32-35, 2006.
- [19] M. Emanoil, S. Alexandra and C. L. Mihaela, "Knowledge workers – drivers to organizational performance in a knowledge-based economy," *Mathematical Methods in Engineering and Economics* , pp. 150-154, 2014.
- [20] M. D'Adamo and A. Kols, "A Tool for Sharing Internal Best Practices," *The INFO Project. United States Agency for International Development (USAID)*, 2005.
- [21] Z. A. Hamid and K. A. M. Kamar, "Modernising the Malaysian Construction Industry," in in *Proceedings: W089 - Special Track 18th CIB World Building Congress (combined with TG71 and TG73)*, Salford, UK , 2010.
- [22] C. Ashton, *Managing Best practices*, London: Business Intelligence, 1998.
- [23] D. Skyrme, "Best Practices in Best Practices. K-Guide. Highclere, England, David Skyrme Associates, June 2002.
- [24] E. Wenger, "Communities of practice: A brief introduction.," 2007.
- [25] M. Massaro, M. Pitts, F. Zanin and R. Bardy, "Knowledge Sharing, Control Mechanisms and Intellectual Liabilities in Knowledge-Intensive Firms," *The Electronic Journal of Knowledge Management*, pp. vol. 12, no. 2, pp. 110–121, 2014.
- [26] M. M. Evans, K. Dalkir and C. Bidian, "A Holistic View of the Knowledge Life Cycle: The Knowledge Management Cycle (KMC) Model," *Electronic Journal of Knowledge Management*, p. Volume 12 Issue 2, 2014 .
- [27] C. Janicot and S. Mignon, "Knowledge codification in audit and consulting firms: a conceptual and empirical approach," *Knowledge Management Research & Practice*, vol. 10, p. 4–15, 2012.
- [28] M. Evans and N. Ali, "Bridging knowledge management life cycle theory and practice," in *Proceedings of the International Conference on Intellectual Capital, Knowledge Management and Organisational Learning – ICICKM*, Washington University, Washington, DC, USA, 2013.
- [29] O. Serrat, "Identifying and Sharing Good Practices," *ADB Knowledge Solutions*, 2008.
- [30] L. Banachowski, AldonaDrabik and J. P. Nowacki, "Improving Process of Teaching Students by Means of Methods and Tools of Knowledge Management and e-Learning," *International Journal of Innovation, Management and Technology*, pp. Vol. 4, No. 6, 599-604, 2013.
- [31] A. Essandoh, "Applying Knowledge Management in Teacher Evaluation Systems," *PhD dissertation, WALDEN UNIVERSITY*, p. 402 pages, 2013.
- [32] M. D. Merrill, "Knowledge Objects," *CBT Solutions*, pp. 1-11, 1998.
- [33] J. Roschelle, M. Koutlis, A. Repenning, J. Phillips, N. Jackiw and D. Suthers, "Developing Educational Software Components," in: *Computer* , pp. pp. 50-58, 1999.
- [34] A. Panar and T. M. Shumaker, "Sharable Content Object Reference Model (SCORM), 4th edition," *Advanced Distributed Learning*, 2009.
- [35] I. Grützner, D. Pfahl and G. Ruhe, "Systematic courseware development using an integrated engineering style method," in *Proceedings of the World Congress "Networked Learning in a Global Environment: Challenges and Solutions for Virtual Education"*, Sliedrecht, 2002.
- [36] N. Friesen, "What are Educational Objects?," *Interactive Learning Environments* , vol. vol. 9, no. no. 3, pp. pp. 219-230, 2001.
- [37] D. A. Wiley, "Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy," In Wiley, D.A. (Ed.), *The Instructional Use of Learning Objects*, 2000.
- [38] K. Cardinaels, E. Duval and H. Olivi', "A Formal Model of Learning Object Metadata," in *Innovative Approaches for Learning and Knowledge Sharing* , vol. Volume 4227, Berlin, Heidelberg, Springer, 2006, pp. pp 74-87.
- [39] B. Melancon, J. Luisi and K. Negyesi, "The Definitive Guide to Drupal 7," *Apress*, 2011.
- [40] N. Gupta, S. Chawla and R. K. Singla, "Proposed System of Automatic Evaluation of Learning Objects," *International Journal of Scientific and Research Publications*, pp. Volume 2, Issue 9, PP. 1-5, September 2012.
- [41] L. Argote and P. Ingram, "Knowledge transfer: A Basis for Competitive Advantage in Firms," *Organizational Behavior and Human Decision Processes*, pp. 82.1: 150-169, 2000.



# An Automated Recommender System for Course Selection

Amer Al-Badarenah

Computer Information Systems Department  
Jordan University of Science and Technology  
Irbid 22110, Jordan

Jamal Alsakran

Computer Science Department  
University of Jordan  
Amman 11942, Jordan

**Abstract**—Most of electronic commerce and knowledge management systems use recommender systems as the underlying tools for identifying a set of items that will be of interest to a certain user. Collaborative recommender systems recommend items based on similarities and dissimilarities among users' preferences. This paper presents a collaborative recommender system that recommends university elective courses to students by exploiting courses that other similar students had taken. The proposed system employs an association rules mining algorithm as an underlying technique to discover patterns between courses. Experiments were conducted with real datasets to assess the overall performance of the proposed approach.

**Keywords**—collaborative recommendation; association rule mining; data mining; recommender system; course selection

## I. INTRODUCTION

In our daily life, we make our choices at most cases relying on recommendations from newspapers, people, or the Internet (e.g., book reviews, movie, restaurant rating, etc.). However, as the amount of information available on the Internet grows, searching for and making decisions about information becomes difficult. New technologies are required to assist Internet users to cope with information overload. Recommender systems [1, 2, 3, 4, 5, 6, 7, 8] have been an important application area and the focus of considerable recent academic and commercial interests. They are widely used by many commercial and non-profit web sites to help users to select items based on users' preferences. The goal of a recommender system is to provide recommendations that users will evaluate favorably and accept [9].

The main steps a recommender system utilizes to propose an item to a user include analyzing user data, extracting useful information, and finally predicting items to users [10]. By and large, recommender systems use combinations of different kinds of user data, including item ratings [11, 12, 35], item features [13, 14], purchase history [15, 16], user demographic data [17, 18], text comments [19, 20], and contextual information [21, 22]. Ideally, a recommender system should allow users to find the preferable items quickly and to alleviate the information overload problem [24, 25]. Currently, a wide range of electronic commerce web sites that sell products, such as books, movies, music, vehicles, and rental apartments, employ online recommendation to fulfill customers' needs [16, 26, 27, 28].

## A. Classifications of Recommender Methods

In general, recommender methods are usually grouped into broad four categories, based on how recommendations are made [4, 15]: collaborative [5, 10, 12, 13, 29, 40, 62, 63], content-based [11, 17, 21, 32, 33, 34], knowledge-based [27, 65, 66, 67, 68, 69], and hybrid recommender approaches [4, 17, 37, 46, 70]. An excellent survey of different recommender systems for various applications can be found in [30]. Pure content-based recommender methods typically propose items to a target user based on affinity between items' contents and the user profile, ignoring data from other users. On the other hand, in pure collaborative recommender methods, items are recommended to a target user based on similarities with other users' preferences (e.g., purchase histories and user ratings), ignoring items' features. Knowledge-based recommender methods exploit inferences, often adopting techniques from artificial intelligence, to deduce a match between user and item [13]. Knowledge-based methods use deep knowledge about features of items rather than users' ratings. To overcome deficiencies associated with pure recommender systems, hybrid approaches based on combining of collaborative recommender method with content-based or knowledge-based approach were proposed [46, 74].

## A. Content-Based Recommender Methods

Content-based recommenders propose items (e.g., articles, music, images) to a target user based on similarities between the content of the yet unseen items and the user's preferences (i.e., items that the user has liked in the past). For example, the system may try to correlate the occurrences of keywords in a web page with the user's interest [36]. A profile of the user's interests is created by analyzing the content of the items that have been rated by the user. Later on, when a user interacts with the recommender system, the items proposed are similar to those items in the user's profile. A variety of techniques from approximation theory, machine learning, and various heuristics have been used for analyzing the items' content and finding regularities in this content that can serve as the basis for making recommendations. Commonly used techniques are nearest neighbor formation [33], Bayesian classifier [38, 39], Neural Networks [40], and Association Rule Mining [41]. Several existing content-based recommender systems have been employed in industry. Among these systems are ACR News [42], InfoFinder [43], and WebSail [44].

A pure content-based recommender system has some drawbacks. Specifically, the content feature extraction methods

are suitable to analyze only textual data [45]. On the other hand, applying content-based systems to non-textual data, say multimedia (e.g., video and audio streams), has remained a difficult task. Such data needs to be manually annotated with content information. A second problem, known as over-specialization [46], is that content-based systems can only propose items that are closely related to the user's profile; therefore, the user will only get to see items that are similar to previous items that the user has indicated interest in. However, in certain cases, the user may be interested in other potential items that are outside his usual preferences. One of the ways to overcome this problem is to filter out high-relevance items if they are too similar to the user profile [47]. Another way to tackle this problem is to use the diversity, in addition to the similarity, for ranking items for recommendation [48]. Finally, a new user with an empty profile has to judge an adequate number of items in order to get a reliable recommendation. This is often referred to as a new user problem, since a new user, having very few ratings, often is not able to get appropriate recommendations [21].

### B. Collaborative Recommender Methods

Collaborative recommender systems recommend items to a target user based on similarity between past preferences of the target user and other similar users. Unlike content-based systems, collaborative systems predict items based on users' ratings and not machine analysis of content. Typically, users are grouped together by employing methods, like clustering [49], based on their preferences. Once the clusters are created, the clusters that have the strongest correlations with the target user are used in making recommendations for a user. Different analytical methods have been used in collaborative recommender systems to represent affinities among users' preferences. Commonly used methods are correlation-based [19], Bayesian network [56], and association rules techniques [57, 58]. Most recently, collaborative recommender systems have gained much attention from academia and industry. Among these systems are GroupLens [19], Ringo [45], the Bellcore Video Recommender [35], the Jester system, which recommends jokes [59], and the PHOAKS system that assists users finding relevant data on the Web [60].

Some of the problems of the content-based systems described before can be fixed using collaborative systems, since collaborative systems do not depend on error-prone machine analysis of the content. The system proposes items to users based on quality of items (i.e., preferences generated by users' ratings), rather than more objective properties of the items themselves (i.e., items' content). This makes the system capable of recommending items to the user, which are very different (content-wise) from what the user has indicated liking before [45]. Unlike content-based systems, a collaborative recommender system can deal with contents of variety of types (e.g. text, artwork, music, video clip); which can enhance the quality of recommendations.

Although collaborative recommender systems have been successfully utilized in both practice and research, they still suffer from some problems including the early rater problem, the sparsity problem, and the gray sheep problem. The early rater problem (also called the cold start problem [53]) refers to the case when recommendations are required for a new item

that has not been yet rated or for a user that just starts to use the system. In this case, the collaborative recommender does not work well, since there is no much rating information on either the item or the user. The sparsity problem, which can also be related to the rater problem [54], occurs when available items ratings are insufficient for identifying similarities between users, leading to poor recommendation. This problem is also known as reduced coverage [26]. The gray sheep problem, on the other hand, occurs when a poor recommendation may be proposed to users with odd preferences compares to the rest of the users, since there will not be any other user who has similar preferences [46].

### C. Knowledge-Based Recommender Methods

Knowledge-based recommendation approaches are especially appropriate for domains where profound product knowledge is needed to recognize and justify solutions; like financial services for instance. In comparing users purchasing simple items like videos or books, with users purchasing financial services, the latter are much more in need of information and of intelligent interaction mechanisms that support appropriate solutions. Therefore, explicit representation of product, marketing, and sales knowledge is needed [12].

Knowledge-based recommender systems use inferences, often adopting techniques from AI, to infer a match between a buyer and a product based on the features of a product rather than the buyers' ratings. Knowledge-based recommender systems are receiving increasing research attention [4, 5, 12, 22, 24, 39]. Many knowledge-based recommender systems that are in existence are an Multimedia enhanced product recommendation [24], an integrated Natural Language Interfaces with personalized recommender system to reduce system-user interactions applied to a restaurant recommender system [39].

### D. Hybrid Recommender Methods

Because of the deficiencies of pure recommender systems, several hybrid recommender systems, that combine collaborative methods with content-based or knowledge based approaches, were proposed. The main goal of a hybrid system is to improve recommendation accuracy as well as to avoid certain drawbacks (e.g., new item and new user problems) of traditional recommender approaches. In his study of hybrid recommender systems [4], Burke describes a taxonomy that consists of seven types of hybrid recommender systems: weighted, switching, mixed, feature combination, cascade, feature augmentation, and meta-level. The study pointed out that most hybrid systems involve the combination of collaborative recommendation with either a content-based or a data mining technique.

This paper presents a novel recommendation methodology that recommends elective courses to a target university student based on affinity between courses taken by the target student and other students. The proposed method is based on collaborative recommender approach that employs association rule mining to discover courses' patterns in order to recommend courses. In this study, clustering is first employed to a group of students based on their courses' grades, then nearest neighborhood is applied to select the students' group



that is most similar to the target student, and finally, association rule mining is used to provide course recommendations to the target student.

The rest of the paper is organized as follows. Section 2 describes background information and related work. Section 3 describes the proposed courses recommendation system. Section 4 presents experimental results with discussion and finally in Section 5 we conclude the paper and suggests possible future work.

## II. RELATED WORK

This section first presents a general formalization of collaborative filtering (CF) and its applicability to course recommendation problem. All the symbols used throughout the paper are summarized in Table 1.

### A. Collaborative Recommendation

Collaborative filtering (CF) techniques have been proven to provide satisfying recommendations to users [35, 45]. CF-based techniques rely on the experiences of similar users, i.e., users who share the same preferences. Specifically, the CF methods recommend the target user new items that have been chosen by those users whose preferences are likely to coincide (at least to some extent) with the target user's preferences. Thus, the goal of a CF algorithm is to predict the rating of an item for a particular user, given the same item's ratings of other users. The nearest-neighbor users are those that exhibit the strongest relevance to the target user. Most of CF-based algorithms use  $k$ -nearest-neighborhood algorithm to recommend items [56]. A typical nearest-neighborhood method uses the following steps in making a recommendation to a user [77].

1) Construct a profile vector  $PV$  for a target user  $ut$  by collecting his ratings of some items.  $PV(ut) = \{R(ut, i) \text{ for some } i \in I\}$

2) Compute the pair-wise similarity  $S(ut, ua)$  between this profile  $PV(ut)$  and the profile of each other user  $PV(ua)$ . The actual definition of similarity function  $S$  depends on the CF algorithm used.

3) Construct a list of target user's nearest neighbors  $NN(ut)$  sorted descending by similarity value and take the  $k$  users most similar to user  $ut$ .  $NN(ut) = Top(S, ut, n, k, minSim)$ , where  $Top$  function returns the top  $k$  similarity values  $S(ut, ua)$  between user  $ut$  and any user  $ua$  such that  $S(ut, ua) \geq minSim$ ,  $a = 1, \dots, n$ , and  $a \neq t$ .

4) Use the nearest neighbor list to calculate a predication rating  $P(ut, i)$  for a new item  $i$  for user  $ut$ . If  $P(ut, i) \geq minPred$ , then recommend item  $i$  to user  $ut$ .

### B. Association Rules Mining

The association rule mining (ARM) has received considerable attention over the last decade. The task of ARM is to find the correlations between items in a dataset by discovering items frequently appeared together in a transactional dataset. Formally, ARM is defined as follows [78].

Let  $I = \{i_1, i_2, \dots, i_m\}$  be a set of items. Let  $D = \{T_1, T_2, \dots, T_n\}$  be a set of transactions, where each transaction  $T$  is a set of items such that  $T \subseteq I$ . An association rule is an implication of the form  $X \Rightarrow Y$ , where  $X \subset I, Y \subset I, X \cap Y = \emptyset$ , and  $X \neq Y \neq \emptyset$ . The rule  $X \Rightarrow Y$  holds in the transaction set  $D$  with confidence  $c$  if  $c\%$  of transactions in  $D$  that contains  $X$  also contain  $Y$ . The rule  $X \Rightarrow Y$  has support  $s$  in the transaction set  $D$  if  $s\%$  of transactions in  $D$  contains  $X \cup Y$ .

Confidence and support are the two fundamental quality heuristics that are used to measure the interestingness of the association rule. The confidence measures the validity of  $X \Rightarrow Y$  as a rule; the lesser the exceptions to the rule, the greater its validity. The support measures the efficiency of the rule. The rules that have confidence and support greater than the user specified minimum confidence and minimum support are called *interesting rules*.

Since the introduction of association rules problem in [78], considerable work has been devoted to design efficient algorithms for mining such rules [79, 80]. These algorithms achieved significant improvements over the previous algorithm and were applicable to large databases. Recent survey of association rules mining can be found in [81].

### C. Recommendations Using Association Rules Mining

This section describes briefly some collaborative recommendation systems that employ association rule mining. Mobasher et al. in [71] proposed a technique for web personalization based on association rule discovery from previous user's transaction data. In this technique, recommendations are produced based on matching the current user session against patterns discovered through association rules on user transaction data. Changchien et al. in [57] presented a method to support on-line recommendation by customers and products fragmentation. It consists of two essential modules; one is clustering based on neural networks, which groups tasks on a large amount of database records, while the other extracts association rules by employing rough set theory.

Additionally, a new personalization recommendation system that integrates both user clustering and association rules techniques was proposed in [72]. The clustering method is used to cluster users' time-framed navigation sessions that are analyzed using association rules technique to establish recommendations for other similar users in the future. In addition, Wang et al. in [73] proposed a personalized recommendation system that incorporates content-based, collaborative, and association rules.

TABLE I. SYMBOLS USED IN THE PAPER

Symbol	Meaning
$U$	The set of users
$n$	The number of users, $ U $
$u, u_1, \dots, u_n$	Users
$I$	The set of items
$m$	The number of items, $ I $
$i, i_1, \dots, i_m$	Items
$S(u_i, u_a)$	The similarity function used
$PV(u)$	The profile vector for a user $u$

R(u, i)	Rating of the user u for the item i
P(u, i)	Prediction of user u's rating for item i
minSim	minimum similarity threshold value
minPred	minimum prediction threshold value

Liu et al. in [58] have proposed a product recommendation methodology that combines group decision-making and association rules. The system addressed the lifetime value of a customer to a firm. This system employed analytical hierarchy process to evaluate customer lifetime value (CLV) then clustering were used to group customers according to the CLV. Finally, an association rule method was implemented to provide product recommendation to each customer group.

#### D. Clustering Techniques

Traditional collaborative filtering techniques are often based on matching the current user profile against clusters of similar profiles obtained by the system over time from other users. Clustering is the process of grouping the data into classes or clusters so that the objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Dissimilarities are assessed based on the attributes values describing the objects. Often distance measures are used. There exist a large number of clustering algorithms in the literature. The choice of clustering algorithms depends both on the type of the data available and on the particular purpose and application. k-mean clustering technique is used in this work because of its simplicity and being suitable to be used with numerical unsupervised data like student courses' grades.

K-means clustering technique is one of the simplest unsupervised learning algorithms that do not rely on predefined classes to solve the clustering problem. The main idea is to define  $k$  centers, one for each cluster. The next step is to take each point belonging to a given data set and associate it to the nearest centers. At this point, we need to re-compute  $k$  new centers of the clusters resulting from the previous step. After we have these  $k$  new centers, a new binding has to be done between the same data set points and the nearest new centers. A loop has been generated. As a result of this loop, we may notice that the  $k$  centers change their location step by step until no more changes are done. In other words, centers do not move any more. Finally, this algorithm aims at minimizing a criterion function, in this case a squared error function. The criterion function  $J$  is defined by

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2$$

where  $\|x_i^{(j)} - c_j\|^2$  is a chosen distance measure between a data point  $x_i^{(j)}$  and the cluster center  $c_j$ , is an indicator of the distance of the  $n$  data points from their respective cluster centers.

#### E. Similarities Techniques

Similarity is quantity that reflects the strength of relationship between two objects or two features. This quantity is usually having range of either -1 to +1 or normalized into 0

to 1. If the similarity between object  $i$  and object  $j$  is denoted by  $S_{ij}$ , we can measure this quantity in several ways depending on the scale of measurement (or data type) that we have.

Similarity distance measures are commonly used for computing the similarity of objects described by interval scaled variables. Interval scale variables are continuous measurements of roughly linear scale. Typical examples include weight, height, grade, and weather temperature. The similarity between the objects described by interval-scaled variables is typically computed based on the distance between each pair of the objects. Different measures, including Manhattan, Euclidean, and Minkowski, can be used. In this paper, we used Manhattan distance.

### III. THE PROPOSED RECOMMENDATION SYSTEM

Recently, many aspects of receiving a college education have been changed. The volume of course-related information available to students is rapidly increasing. This abundance of information has created the need to help students find, organize, and use resources that match their individual goals, interests, and current knowledge. One of the concerns students have is to make decisions about which courses to take. The concern is more serious for graduate students who have more freedom to choose courses while they care more about taking courses that contribute to their progress towards career goals. To make these decisions, they use information from course catalogs and schedules, consult with their advisors, and seek guidance from their classmates, especially those with similar interests. To give better decision making support to students who wish to make relevant course choices, we have developed a course recommendation system that recommends courses to students based on other similar students.

Our collaborative recommendation system tries to recommend elective courses to students based on what other similar students have taken. It recommends courses and specifies expected grades for these courses. Accordingly, the student may take a course that is recommended by the system with an acceptable grade. Typically, students have the choice to take courses from a set of elective courses and in most cases, the students take the advices from other students that took such courses. In our recommendation system, we automatically find similar students and then apply association rule mining algorithm on their courses to create courses association rules. Discovered courses association rules are used to get recommendation.

To obtain courses association rules, courses dataset is built by mapping each course either compulsory or elective to an item and each student to a transaction. For each student, a transaction is created that contains the grades of all courses taken by the student. Table 2 shows an example of courses dataset, where -1 indicates that the student does not take the corresponding course. Examples of courses association rules that can be generated are "70% of students who got **A** in **Algo** course and **E** in **Parallel** course may also get **B** in **OS** course" and "90% of students who got **D** in **Net** course may also get **D** in **Mining** course".

The system takes as an input the specified minimum support, the specified minimum confidence, and the courses

dataset. As an output, the system generates courses association rules that satisfy the support and confidence constraints. Then the system uses these rules to generate courses recommendations. Fig. 1 illustrates our recommendation methodology. The main steps of our system are described in the following subsections.

TABLE II. AN EXAMPLE OF COURSES DATASET

Student ID	Alg	Mining	OS	Parallel	Net
1000	A	A	-1	F	B
1001	F	F	C	D	E
1002	B	A	F	E	E
1003	E	D	B	-1	A

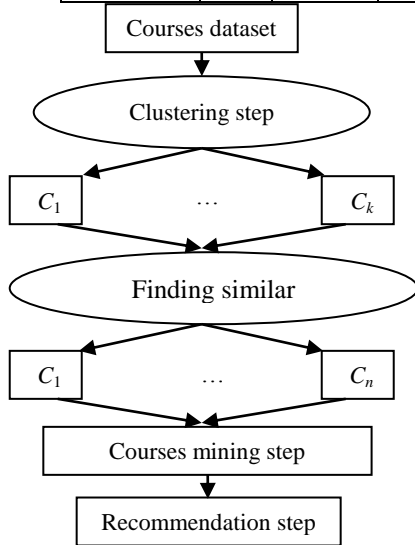


Fig. 1. Proposed recommendation methodology

### A. Clustering Step

This step applies clustering technique on courses dataset to group similar students to the same cluster. The clustering algorithm used in this step is *k-means* clustering algorithm [21], where each cluster is represented by the mean value of students in the cluster. A cluster of similar students is treated as one group in this step. The similarity between the courses is typically computed based on the distance between each pair of courses. A well-known distance measure is *Manhattan* distance [21], which measures the distance between two  $p$ -dimensional data objects  $i = (x_{i1}, x_{i2}, \dots, x_{ip})$  and  $j = (x_{j1}, x_{j2}, \dots, x_{jp})$  as follow  $d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$ . In our case,  $i$  and  $j$  represent two students to calculate similarity between them,  $x_{i1}, x_{i2}, \dots, x_{ip}$  represent the courses' grades for student  $i$ , while  $x_{j1}, x_{j2}, \dots, x_{jp}$  represent the courses' grades for student  $j$ . For example, *Manhattan* distance between the first two students, 1000 and 1001 of Table 2 is:

$$d(1000, 1001) = |A - F| + |A - F| + |F - D| + |B - E| = 15$$

where the grades (A, B, C, D, E, and F) are mapped to the integers (1, 2, 3, 4, 5, and 6), respectively. Here, we did not include the distance between students' grades in OS course because the first student did not take this course.

### B. Finding Similar Students

In this step, the group most similar to the target student is selected by comparing his/her previous courses' grades with the courses' grades of the mean student of the  $n$ -cluster. We used *n-nearest neighborhood* technique [22] to select the most  $n$  similar groups generated in Step I. In this way, all of the clusters centers are represented in a  $p$ -dimensional pattern space. When given an unknown student, an *n-nearest neighborhood* searches the pattern space for the  $n$  clusters that are closest to the unknown student. These  $n$  clusters are the  $n$  "nearest neighbors" of the unknown sample. Closeness is defined in terms of Manhattan distance.

### C. Courses Mining

Association rules mining is used to discover courses association rules. The algorithm is applied on the  $n$  similar groups created from Step II. Table 3 shows an example of the transactional courses dataset that is used in the mining step, where each transaction contains transaction Id (*TID*) that corresponds to student *ID* and set of items. Items are in the form of  $[course:grade]$  pairs. The rationale of using  $[course:grade]$  pairs is that our system tries, in addition to recommend courses, also to specify expected grades for those courses. This will help students to select elective courses with high grade expectation. An example of courses association rule is  $[Alg:A] \wedge [OS:D] \Rightarrow [Parallel:C]$ , this indicates that if the target student got A in **Algo** and D in **OS** courses, then the system recommends the student to take **Parallel** course with expected grade C.

### D. Recommending courses

The courses association rules, generated in Step III, are used to recommend elective courses. This step is the core step in our recommendation system. The recommendation is done after interesting courses association rules are created. The association rules are in the form of  $[c_1:g_1] \wedge \dots \wedge [c_n:g_n] \Rightarrow [c_{n+1}:g_{n+1}] \wedge \dots \wedge [c_m:g_m]$ , where  $[c:g]$  represents a course  $c$  with its grade  $g$ . The whole recommendation strategy is described as follows:

1) If the target student has taken the courses  $[c_1, \dots, c_n]$ , the ones in the antecedent part of the rule, with grades  $[g_1, \dots, g_n]$  respectively, then the system recommends the student to take the elective courses  $[c_{n+1}, \dots, c_m]$ , the ones in the consequent part of the rule, with expected grades  $[g_{n+1}, \dots, g_m]$ , respectively. The system may also recommend courses from the consequent part of the rule if they met specified minimum grades. In this step, the student can specify a minimum grade in which the system will not recommend any course with grade that does not exceed this minimum threshold. In the following example, the system recommends courses for a student who has taken the following courses  $[DB:A]$ ,  $[Parallel:C]$ ,  $[Net:D]$ , and specified the minimum grade to C.

Rule#1:  $[DB:A] \wedge [Parallel:C] \Rightarrow [Mining:B]$

Rule#2:  $[DB:A] \wedge [Net:D] \Rightarrow [AI:B] \wedge [Archit:D]$

1. Rule#1 recommends **[Mining:B]**
2. Rule#2 recommends **[AI:B]**

TABLE III. AN EXAMPLE OF TRANSACTIONAL COURSES DATASET USED IN MINING STEP

TID	Items
1000	[Alg:A], [Mining:A], [Parallel:F], [Net:B]
1001	[Alg:F], [Mining:F], [OS:C], [Parallel:D], [Net:E]
1002	[Alg:B], [Mining:A], [OS:F], [Parallel:E], [Net:E]
1003	[Alg:E], [Mining:D], [OS:B], [Net:A]

Because the student has taken the courses in the antecedent part of rule#1, it recommends the student the **Mining** course with expected grade is **B**. In the same way rules#2 recommends the student the **AI** course with expected grade is **B**. The rule#2 does not recommend the student to take the **Archit** course because the expected grade **D** is less than the minimum threshold already specified by that student.

2) The target student might not take all the courses in the antecedent part of the rule. The rule still can be used in recommendation and a new constraint called match is used to assess the quality of rule in the recommendation. The match is defined as the number of matched courses between the target student's courses and the courses in the antecedent part of the rule.  $match_{rule} = \frac{\text{matched courses between student and antecedent part of rule}}{\text{total courses in antecedent part of rule}}$

If  $match_{rule}$  value is greater than a threshold value, then the rule is used for recommendation.

3) If different rules recommend the target student different courses, then the student can select the top-N courses recommended by best quality rules, i.e., rules that have the highest supports, the highest confidences, and the highest match. If there a tie between rules, the system recommends courses in this order of preferences: highest support rule, highest confidence, and highest match.

4) The recommendation system does not recommend new students since they do not have courses taken yet.

For example, consider the following four course association rules.

- Rule#1:  $[DB:A] \wedge [Parallel:C] \Rightarrow [Mining:B]$
- Rule#2:  $[DB:A] \wedge [Net:D] \Rightarrow [AI:B] \wedge [Archit:D]$
- Rule#3:  $[Mining:E] \wedge [OS:C] \Rightarrow [IR:D]$
- Rule#4:  $[Net:D] \wedge [Parallel:C] \Rightarrow [Image:B]$

Table 4 shows the proposed recommendations, that are derived from the above listed four rules, for a student who took the following courses: **[DB:A]**, **[Parallel:C]**, **[Net:D]**, with **C** as a minimum accepted grade and 50% as a match threshold value. As shown in Table 4, Rule#1 has  $match_{rule}$  value equal to 66%, which is above the match threshold 50%. Then this rule recommends the elective course in the consequent part **Mining** with **B** as expected grade. Rule#2 has also  $match_{rule}$  value equals to 66%. The elective courses in the consequent part of the rule are **[AI:B]** and **[Archit:D]**. The rule recommends only **[AI:B]**, but not **Archit** course, because the expected grade **D** is less than the specified minimum accepted grade **C**. Rule#3 is not used in recommendation because its  $match_{rule}$  value (zero value) is less than 50%, the match

threshold value. The last rule Rule#4 recommends the elective course **[Image:B]**.

TABLE IV. THE PROPOSED RECOMMENDATIONS USING THE GIVEN FOUR COURSES ASSOCIATION RULES

Rule used	Recommendation	$match_{rule}$
#1	<b>[Mining:B]</b>	66%
#2	<b>[AI:B]</b>	66%
#2	<b>[Archit:D]</b> is not recommended since $D < C$	66%
#3	No recommendation since $match_{rule} < 50\%$	0%
#4	<b>[Image:B]</b>	66%

Another example consists of eight course association rules.

Student courses: **[DB:A]**, **[Parallel:C]**, and **[Net:D]**

Rule#1:  $[DB:A] \wedge [Parallel:C] \Rightarrow [Mining:B]$

Rule#2:  $[DB:A] \wedge [Net:D] \Rightarrow [AI:B] \wedge [Archit:D]$

Rule#3:  $[Mining:E] \wedge [OS:C] \Rightarrow [IR:D]$

Rule#4:  $[Net:D] \wedge [Parallel:C] \wedge [Net:D] \Rightarrow [Image:B]$

Rule#5:  $[DB:A] \wedge [Parallel:D] \wedge [IR:D] \Rightarrow [Alg:B]$

Rule#6:  $[DB:B] \wedge [Net:D] \wedge [OS:C] \Rightarrow [AI:B] \wedge [Archit:D]$

Rule#7:  $[Alg:E] \wedge [OS:C] \wedge [DB:A] \wedge [Image:B] \Rightarrow [IR:D]$

Rule#8:  $[Net:D] \Rightarrow [Artchit:B] \wedge [Image:B]$

Table 5 shows the proposed recommendations, that are derived from the above listed eight rules, for a student who took the following courses: **[DB:A]**, **[Parallel:C]**, **[Net:D]**, with **C** as a minimum accepted grade and 50% as a match threshold value. As shown in Table 5, Rule#1 has  $match_{rule}$  value equal to 66%, which is above the match threshold 50%. Accordingly, this rule recommends the elective course in the consequent part **Mining** with **B** as expected grade. Rule#2 has also  $match_{rule}$  value equals to 66%. The elective courses in the consequent part of the rule are **[AI:B]** and **[Archit:D]**. The rule recommends only **[AI:B]**, it does not recommend **Archit** course, because the expected grade **D** is less than the specified minimum accepted grade **C**. Rules #3, #5, #6, and #7 are not used in recommendation because their  $match_{rule}$  values 0%, 33%, 0%, and 25% respectively, are less than 50%, the match threshold value. **[Image:B]** is recommended by rule#4 that has  $match_{rule}$  value 100%. Two courses are recommended using rule#8, **[Artchit:B]** and **[Image:B]**.

TABLE V. THE PROPOSED RECOMMENDATIONS USING THE GIVEN EIGHT COURSE ASSOCIATION RULES

Rule used	Recommendation	$match_{rule}$
#1	<b>[Mining:B]</b>	66%
#2	<b>[AI:B]</b>	66%
#2	<b>[Archit:D]</b> is not recommended since $D < C$	66%
#3	No recommendation since $match_{rule} < 50\%$	0%
#4	<b>[Image:B]</b>	100%
#5	No recommendation since $match_{rule} < 50\%$	33%
#6	No recommendation since $match_{rule} < 50\%$	0%
#7	No recommendation since $match_{rule} < 50\%$	25%
#8	<b>[Artchit:B]</b> and <b>[Image:B]</b>	100%

#### IV. EXPERIMENTS

We performed experiments using courses dataset taken by 2000 graduate students of Electrical Engineering. The total

number of courses is 54, where five of them are compulsory, while the remaining are elective courses. Each student has a choice of selecting four elective courses. As early mentioned, *k*-means clustering algorithm is used, with *k* = 5. In the mining step, we used *n* = 3 as the most similar *n* clusters. We divided our dataset into two disjoint sets: the *training set* and the *test set*. *K*-fold cross validation technique is applied on the dataset, in which we run *k* experiments, each time setting aside different 1/*k* of the data to test on, and average the result. Popular values for *k* are 5 and 10, we use *k* = 5. Mining association rules algorithm is applied on training set to generate courses association rules. Then we measured the percentage of students in the test set that were correctly recommended by courses association rules.

To evaluate the performance of our recommendation system, we use two standard information retrieval measures, *precision* and *recall* [24]. Precision is the percentage of the number of recommended courses taken to the total number of recommended courses, while recall is the percentage of the number of recommended courses taken to the total number of courses taken by the students. More precisely

$$Precision = \frac{\# \text{ of recommended courses taken}}{\text{total \# of recommended courses}}$$

$$Recall = \frac{\# \text{ of recommended courses taken}}{\text{total \# of courses taken by students}}$$

For recommendation systems, precision is more significant than recall, because we concern more on getting high quality recommendation than just recommending a large number of courses. Therefore, our goal is to achieve a high precision with reasonably high recall. The main parameters used in our experiments are minimum confidence, minimum match, minimum specified grade, and minimum support. The following sections show experiments that were performed in order to choose the appropriate values for the parameters.

#### A. Minimum Confidence

Both minimum support and minimum confidence are important factors that influence the performance of the whole process of recommendation system. Since minimum confidence is used during recommendation step, it would be interested to study the performance of our recommendation system using different minimum confidence values. The results are shown in Fig. 2. As shown in Fig. 2, the minimum confidence value has a significant impact on the performance, i.e., the higher the minimum confidence, the higher the precision but the lower the recall. We achieved the highest precision of 0.95 with a recall of 0.5 for the minimum confidence of 90%. Even though we think that the precision is the most important factor in recommendation systems, the best combination of precision and recall values, which occurs with minimum confidence of 80%, is also important in the sense that we can achieve higher values of both precision and recall.

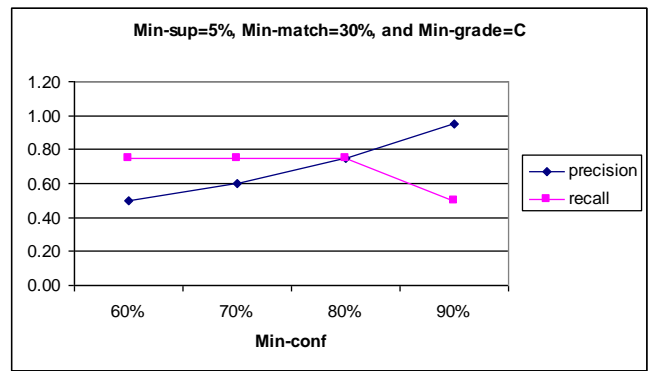


Fig. 2. Performance using different minimum confidence values

#### B. Minimum Match

In order to select the appropriate minimum match, we performed some experiments using different minimum match values. As shown in Fig. 3, the higher the minimum match the higher precision but the lower the recall. We achieved the highest precision with minimum match 90%. When the minimum match varies, a tradeoff between precision and recall values is noticed.

#### C. Minimum Grade

Minimum specified grade is also an important parameter of our approach. Fig. 4 gives the performance for different minimum grades. We could see a general fact that the minimum grade has a similar influence on the performance as the minimum confidence, i.e., the higher the minimum grade, the higher the precision but the lower the recall. When the minimum grade is varied, there shows a tradeoff between the precision and the recall.

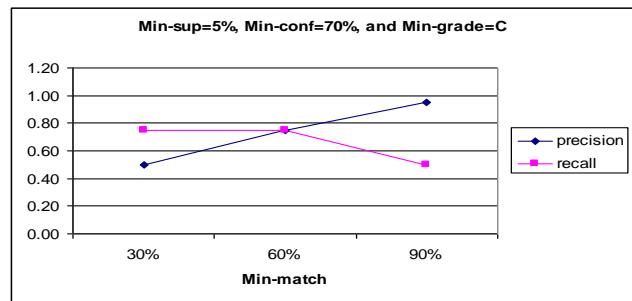


Fig. 3. Performance using different minimum match values

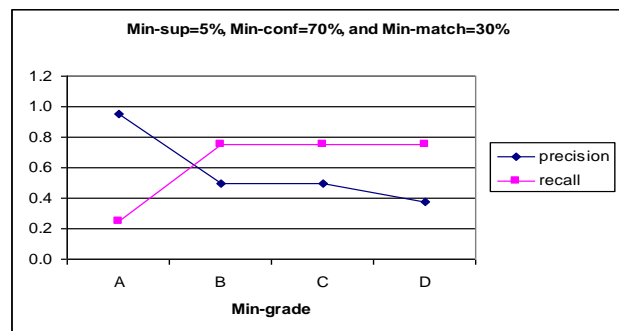


Fig. 4. Performance using different minimum grade values



#### D. Minimum Support

In our experiments for courses associations, we tested the performance for different minimum supports, i.e., we only use rules above a specified minimum support for recommendation during a test. The results are shown in Fig. 5.

Form our observations, we found that when a target student's minimum support determined by the mining process is very low, it takes a very long time to mine rules for this student and at the same time the performance is bad. While, if a student's minimum support is greater than a threshold, then we get a better performance.

#### V. CONCLUSION AND FUTURE WORK

The main contribution of this paper is a new collaborative recommendation system that employed association rules algorithm to recommend university elective courses to a target student based on what other similar students have taken. The experiments shown that association rule is a desirable tool for making recommendation to a target student. Through our experiments, we noticed the patterns of influence of different parameters on the performance of the system. The confidence and match of a rule have a great impact on the performance, but the highest confidence or match may not be the best choice. By choosing a relatively high confidence or match, we can achieve a better performance.

Much work can be performed in the future such as doing comparison between our method and other typical methods. In addition, further experimental evaluation, joining collaborative and content-based recommendations, and applying the new recommendation system in other domains of interest is expected as future work.

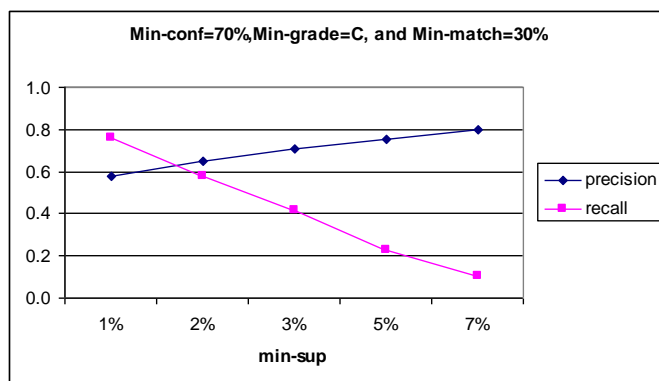


Fig. 5. Performance using different minimum support values

#### REFERENCES

- [1] Resnick P. and Varian H., Recommender Systems, *Communication of the ACM*, 40(3), 56-58, March 1997.
- [2] Montaner M., Lopez B., and Josep R., A Taxonomy of Recommender Agents on the Internet, *Artificial Intelligence Review*, 19, 285-330, 2003.
- [3] Bridge D., Product recommendation systems: A new direction. In R.Weber, Wangenheim, C., eds.: *Procs. of the Workshop Programme at the Fourth International Conference on Case-Based Reasoning*. Vancouver, Canada, 79-86, 2001.
- [4] Burke R., Hybrid Recommender Systems: Survey and Experiments, *User Modelling and User-Adapted. Interaction* 12, 4, 331-370, 2002.

- [5] Herlocker J., Konstan J., Tervin L., and Riedl J., Evaluating Collaborative Filtering Recommender Systems, *ACM Transactions on Information Systems*, 22(1), 5-53, 2004.
- [6] Mirza J., Keller B., and Ramakrishnan N., Studying Recommendation Algorithms by Graph Analysis, *Journal of Intelligent Information Systems*, 20(2), 131-160, 2003.
- [7] Ariely D., Lynch, J., and Aparicio M., Learning by Collaborative and Individual-Based Recommendation Agents. *Journal of Consumer Psychology*, 14(1 & 2), 81-95, 2004.
- [8] Ziegler C., McNee S., Konstan J., and Lauen G., Improving Recommendation Lists Through Topic Diversification. In *Proceedings of 14th Int. WWW Conference*, Chiba, Japan, 22-32, 2005.
- [9] Gretzel U. and Fesenmaier D., Persuasion in Recommender Systems, *International Journal of Electronic Commerce*, 11(2), 81-100, 2006.
- [10] Chen A. and McLeod D. (2006). Collaborative Filtering for Information Recommendation Systems, *Encyclopedia of E-Commerce, E-Government, and Mobile Commerce*. IGI Global, pp. 118-123.
- [11] Melville P., Mooney R., and Nagarajan R., Content-Boosted Collaborative Filtering for Improved Recommendations, In *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-2002)*, (July 2002), Edmonton, Canada, 187-192.
- [12] McLaughlin M. and Herlocher J., A Collaborative Filtering Algorithm and Evaluation Metric that Accurately Model the User Experience, In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. (2004), Sheffield, United Kingdom, 329-336.
- [13] Ahn H. J., Utilizing Popularity Characteristics for Product Recommendation. *International Journal of Electronic Commerce*, (2006), 11(2), 59-80.
- [14] Nikolaeva R. and Sriram S., The Moderating Role of Consumer and Product Characteristics on the Value of Customized Online Recommendations, *International Journal of Electronic Commerce*. (2006), 11(2), 101-124.
- [15] Huang Z., Chung W., and Chen H., A Graph Model for E-Commerce Recommender Systems, *Journal of the American Society for Information Science and Technology (JASIST)*, (2004), 55(3), 259-274.
- [16] Schafer J., Konstan J., Riedl J., E-Commerce Recommendation Applications, *Data Mining and Knowledge Discovery*, (2001), 5(1-2), 115-153.
- [17] Pazzani M., A Framework for Collaborative, Content-Based and Demographic Filtering, *Artificial Intelligence Review*, (1999), 13(5-6), 393-408.
- [18] Chen M., Chiu A., and Chang H., Mining Changes in Customer Behavior in Retail Marketing. *Expert Systems with Applications*, (2005), 28(4), 773-781.
- [19] Resnick P., Iacovou N., Suchak M., Bergstrom P., and Riedl J., GroupLens: An Open Architecture for Collaborative Filtering of Netnews, In *Proceedings of ACM Conference on Computer Supported Cooperative Work*, (1994), Chapel Hill, NC, USA, 175-186.
- [20] Goldberg D., Nichols D., Oki B., and Terry D., Using Collaborative Filtering to Weave an Information Tapestry, *Communications of the ACM*, (1992), 35(12), 61-70.
- [21] Adomavicius G., Sankaranarayanan R., Sen S., and Tuzhilin A., Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach, *ACM Transactions on Information Systems*, (2005), 23(1), 103-145
- [22] Yap G., Tan A., and Pang H., Dynamically Optimized Context in Recommender Systems, In *the Proceedings of the 6th International Conference on Mobile Data Management*, (May 2005), Ayia Napa, Cyprus, 265-272.
- [23] Good N., Scafer J., Konstan J., Borchers AI, Sarwar B., Herlocaker J., and Riedl J., Combining Collaborative Filtering with Personal Agents



- for Better Recommendations, *In Proceedings of the Eleventh Innovative Applications of Artificial Intelligence Conference*, July 18-22, 1999, American Association for Artificial Intelligence, Menlo Park, CA, USA, 439-446.
- [24] Maes P., Agent that Reduce Work and Information Overload, *Communication of the ACM*, (1994), 37(7), 31-40.
- [25] Sarwar B., Karypis G., Konstan J., and Riedl J., Analysis of Recommendation Algorithms for E-Commerce. *In Proceedings of the 2<sup>nd</sup> ACM Conference on Electronic Commerce (EC-00)*, October 17-20, 2000, Minneapolis, MN, USA, 158-167.
- [26] Burke R., Knowledge-Based Recommender Systems, In A. Kent (ed.) *Encyclopedia of Library and Information Systems*, (2000), vol 69(32).
- [27] Zhang J. and Pu P., A Comparative Study of Compound Critique Generation in Conversational Recommender Systems. *In Proceedings of the 4<sup>th</sup> International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2006)*, June 2006, Dublin, Ireland, Volume 4018 of Lecture Notes in Computer Science, 234-243, Springer.
- [28] Wiranto, Winarko E., Hartati S., and Wardoyo R., Improving the Prediction Accuracy of Multicriteria Collaborative Filtering by Combination Algorithms, *International Journal of Advanced Computer Science and Applications*, 5(4), pp. 52-58, 2014.
- [29] Adomavicius G. and Tuzhilin A., Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, (June 2005), 17(6), 734-749.
- [30] Lee T., Chun J., Shim J., and Lee S., An Ontology-Based Product Recommender System for B2B Marketplaces, *International Journal of Electronic Commerce*, (Jan 2007), 11(2), 125-154.
- [31] Jian C., Jian Y., and Jin H., Automatic Content-Based Recommendation in e-Commerce, *In the proceedings of IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'05)*, (March 29-April 01 2005), Hong Kong, China, 748-753.
- [32] Philip S., Shola P.B., John A.O., Application of Content-Based Approach in Research Paper Recommendation System for a Digital Library, *International Journal of Advanced Computer Science and Applications*, 5(10), pp. 37-40, 2014.
- [33] Hill W., Stead L., Rosenstein M., and Furnas G., Recommending and Evaluating Choices in a Virtual Community of Use, *In the Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, (May 07-11 1995), Denver, Colorado, United States, 194-201.
- [34] Pazzani M., Muramatsu J., and Billsus D., Syskill & Webert: Identifying Interesting Web Sites, *In Proceedings of the National Conference on Artificial Intelligence (AAAI96)*, (1996), Portland, 54-61.
- [35] Shih Y. and Liu D., Hybrid Recommendation Approaches: Collaborative Filtering via Valuable Content Information, *In the Proceedings of the 38<sup>th</sup> Annual Hawaii International Conference HICSS '05*, 2005, 217-223.
- [36] Mooney R. and Roy L., Content-Based Book Recommending Using Learning for Text Categorization, *In Proceedings of the 5<sup>th</sup> ACM Conference on Digital Libraries*, (June 2-7 2000), San Antonio, Texas, USA, 195-204.
- [37] Condliff M., Lewis D., Madigan D., and Posse C., Bayesian Mixed-Effects Models for Recommender Systems, *In ACM SIGIR'99 Workshop on Recommender Systems: Algorithms and Evaluation*, 1999, 23-30.
- [38] Kim E., Kim M., and Ryu J., Collaborative Filtering Based on Neural Networks Using Similarity, *In the Proceedings of Computational Science and Its Applications conference (ICCSA 2)*, (May 2005), Singapore, 355-360.
- [39] Smyth B., McCarthy K., Reilly J., O'Sullivan D., and McGinty L., David C. Wilson, Case-Studies in Association Rule Mining for Recommender Systems, *In Proceedings of the 2005 International Conference on Artificial Intelligence, ICAI 2005*, (June 27-30 2005), Las Vegas, Nevada, USA, 809-815.
- [40] Mobasher B., Cooley R., and Srivastava J., Automatic Personalization Based on Web Usage Mining, *Communications of the ACM*, (2000), 43(8), 142-151.
- [41] Krulwich B. and Burkley C., The InfoFinder Agent: Learning User Interests through Heuristic Phrase Extraction, *IEEE Expert: Intelligent Systems and Their Applications*, (1997), 12(5), 22-27.
- [42] Chen Z, Meng X, Zhu B., and Fowler R., WebSail: From On-line Learning to Web Search, *In Proceedings of the 2000 International Conference on Web Information Systems Engineering*, (2000), Hong Kong, China, 206-213.
- [43] U. Shardanaand P. and Maes P., Social Information Filtering: Algorithms for Automating "Word of Mouth" *Proceedings of the SIGCHI conference on Human factors in computing systems*, (1995), Denver, Colorado, United States, 210-217.
- [44] Balabanovic M. and Shoham Y., Fab: Content-Based, Collaborative Recommendation, *Communications of the ACM*, (1997), 40(3), 66-72.
- [45] Billsus D. and Pazzani M., User Modeling for Adaptive News Access, *Journal of User Modeling and User-Adapted Interaction*, (2000), 10(2-3), 147-180.
- [46] Smyth B. and McClave P., Similarity vs. Diversity, *In Proceedings of the 4<sup>th</sup> International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development*, (August 02, 2001), Vancouver, British Columbia, Canada p, 347-361.
- [47] Kanungo T., Mount D., Netanyahu N., Piatko C., Silverman R., and Wu A., An Efficient k-Means Clustering Algorithm: Analysis and Implementation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 881-892, 2002.
- [48] Schein A., Popescul A., Ungar L., and Pennock D., Methods and Metrics for Cold-Start Recommendations. *In Proceedings of the 25<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2002), Tampere, Finland , 253-260
- [49] Papagelis M., Plexousakis D., and Kutsuras T., Alleviating the Sparsity Problem of Collaborative Filtering Using Trust Inferences, *In Proceedings of iTrust 2005*, (2005), Paris, France, 224-239.
- [50] Breese J., Heckerman D., and Kadie C., Empirical Analysis of Predictive Algorithms for Collaborative Filtering, *In Proceedings of the 14<sup>th</sup> Conference on Uncertainty in Artificial Intelligence*, (1998), Madison, WI, 43-52.
- [51] Changchien S. and Lu T., Mining Association Rule Procedure to Support On-Line Recommendation by Customers and Products Fragmentation, *Expert Systems with Application* , (2001), 20, 325-335.
- [52] Liu D. and Shih Y., Integrating AHP and Data Mining for Product Recommendation Based on Customer Lifetime Value, *Information and Management*, (March 2005), 42(3), 387-400.
- [53] Goldberg K, Roeder T, Gupta D., and Perkins C., Eigentaste: A Constant Time Collaborative Filtering Algorithm, *Information Retrieval*, (July 2001), 4(2), 133-151.
- [54] Terveen L., Hill W., Amento B, McDonald D., and Creter J., PHOAKS: A System for Sharing Recommendations, *Communications of the ACM*, (March 1997), 40(3), 59-62.
- [55] Greco G., Greco S., and Zumpano E., Collaborative Filtering Supporting Web Site Navigation, *AI Communications*, (2004), 17(3), 155-166.
- [56] Konstan J., Miller B., Maltz D., Herlocker J., Gordon L., and Riedl J., GroupLens: Applying Collaborative Filtering to Usenet News, *Communications of the ACM*, (March 1997), 40(3), 77-87.
- [57] Lee D., Kim G., and Choi H., A Web-Based Collaborative Filtering System, *Pattern Recognition*, (2003), 36, 519-526.
- [58] Burke R., Hammond K., and Young B., The FindMe Approach to Assisted Browsing, *IEEE Expert: Intelligent Systems and Their Applications*, (July 1997), 12(4), 32-40.

- [59] Felfernig A. and Gula B., An Empirical Study on Consumer Behavior in the Interaction with Knowledge-Based Recommender Applications, In the Proceedings of the 8<sup>th</sup> IEEE International Conference on E-Commerce Technology and the 3<sup>rd</sup> IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06), 2006, 37-45.
- [60] Felfernig A., Friedrich G., Jannach D., and Zanker M., An Integrated Environment for the Development of Knowledge-Based Recommender Applications, *International Journal of Electronic Commerce*, (2006), 11(2), 11-34
- [61] Felfernig, A., A Knowledge-Based Interactive Selling of Financial Services with FSAdvisor. In N. Jacobstein and B. Porter (eds.), *Seventeenth Innovative Applications of Artificial Intelligence Conference*. (2005), Menlo Park, CA, 1475-1482.
- [62] Towle B. and Quinn C., Knowledge Based Recommender Systems Using Explicit User Models. In *Knowledge-Based Electronic Markets, Papers from the AAAI Workshop, AAAI Technical Report WS-00-04*. (2000), Menlo Park, CA, 74-77.
- [63] Wakil K., Bakhtyar R., Ali K., and Alaadin K., Improving Web Movie Recommender System Based on Emotions, *International Journal of Advanced Computer Science and Applications*, 6(2), 2015, 218-226.
- [64] Mobasher B., Dai H., Luo T., and Nakagawa M., Effective Personalization Based on Association Rule Discovery from Web Usage Data, In the Proceedings of the 3<sup>rd</sup> International Workshop on Web Information and Data Management, (November 09-01, 2001), Atlanta, Georgia, USA, 9-15.
- [65] Wang F. and Shao H., Effective Personalized Recommendation Based on Time-Framed Navigation Clustering and Association Mining. *Expert Systems with Application*, (2004), 27(3), 365-377.
- [66] Wang Y., Chuang Y., Hsu M., Keh H., A Personalized Recommender System for the Cosmetic Business. *Expert Systems with Application*, (2004), 26, 427-434.
- [67] Burke R., Integrating Knowledge-Based and Collaborative-Filtering Recommender Systems, In *AAAI Workshop on AI in Electronic Commerce*, 1999, 69-72.
- [68] Chen Y., Cheng L., and Chuang C., A Group Recommendation System with Consideration of Interactions Among Group Members, *Expert Systems with Applications* (2008), 34(3), 2082-2090.
- [69] Agrawal R., Imilienski T., Swami A., Mining Association Rules Between Sets of Items in Large Databases, In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1993, Washington, DC 207-216.
- [70] Agrawal R. and Srikant R., Fast Algorithms for Mining Association Rules, In Proceedings of the 20<sup>th</sup> International Conference on Very Large Databases, Santiago, Chile: Sept. 1994, 487-499.
- [71] Tsay YJ and Chiang JY, CBAR: An Efficient Method for Mining Association Rules. *Knowledge Based Systems*, 2004, 1-7.
- [72] Ceglar A. and Roddick J., Association Mining, *ACM Surveys*, 2006, V 38(2), article No. 5.

# Fault Tolerant System for Sparse Traffic Grooming in Optical WDM Mesh Networks Using Combiner Queue

Sandip R. Shinde

Research Scholar, Sathyabama University, Chennai &  
Assistant Professor, Vishwakarma Institute of Technology,  
Pune, India

Dr. S. Emalda Roslin

Associate Professor, Department of Electronics &  
Communication Engineering, Sathyabama University,  
Chennai, India

Dr. Suhas H. Patil

Professor, Dept of Computer Engineering, College of  
Engineering, Bharati Vidyapeeth University,  
Pune, India

Archana S. Shinde

Assistant Professor, Department of Information Technology,  
Sinhagad Acedamy Of Engineering,  
Pune, India

**Abstract**—Queuing theory is an important concept in current internet technology. As the requirement of bandwidth goes on increasing it is necessary to use optical communication for transfer of data. Optical communication at backbone network requires various devices for traffic grooming. The cost of these devices is very high which leads to increase in the cost of network. One of the solutions to this problem is to have sparse traffic grooming in optical WDM mesh network. Sparse traffic grooming allows only few nodes in the network as grooming node (G-node). These G-nodes has the grooming capability and other nodes are simple nodes where traffic grooming is not possible. The grooming nodes are the special nodes which has high cost. The possibility of faults at such a node, or link failure is high. Resolving such faults and providing efficient network is very important. So we have importance of such survivable sparse traffic grooming network.

Queuing theory helps to improve the result of network and groom the traffic in the network. The paper focuses on the improvement in performance of the backbone network and reduction in blocking probability. To achieve the goals of the work we have simulated the model. The main contribution is to use survivability on the sparse grooming network and use of combiner queues at each node. It has observed that Combiner queuing alone does the job of minimizing blocking probability and balancing the load over the network. The model is not only cost effective but also increases the performance of network and minimizes the call blocking probability.

**Keywords**—optical communication; sparse traffic grooming; survivability; fault tolerance; Combiner Queue; WDM

## I. INTRODUCTION

Due to increasing demand of bandwidth it is very important for service providers to provide efficient and cost effective solutions to fulfill the need of society. Bandwidth requirements are fulfilled by the use of optical fibers. Optical fibers are providing terabytes of bandwidth. The signal attenuation and distortion is low and also there are very less power requirements. Due to these factors optical network is widely used in the data network. WDM (Wavelength Division

Multiplexing) is used to fulfill the requirement of users over optical network. The main idea of optical WDM network is that, maximum number of users can use network for the transfer of data. Our network must have sufficient bandwidth to fulfill the requirement of bandwidth hungry society. Concurrency will be provided in transmission among the various users in optical communication. WDM divides the bandwidth into multiple frequencies or channels. These are non-overlapping wavelength channels. It is very important to provide support for low traffic rate than the available capacity of wavelength. A single optical channel of STS-1 (51.84Mbps) may be used for that purpose which provides support up to wavelength capacity.

It is very important to transfer low speed data traffic on high-speed networks. Traffic grooming will transfer this low speed traffic to high-speed. The capacity of each wavelength is OC-n it means  $n \times 51.84\text{Mbps}$ . In traffic grooming you have to establish a light path to fulfill the requirement of the end user. The light path is established from source to destination from very low capacity to high capacity up to the maximum capacity of the channel. This traffic grooming problem is subcategories into four problems

- Design of virtual topology with light paths
- Routing the light paths over the physical topology
- Assignment of wavelengths to light paths
- Traffic routing on the virtual topology

For providing these services at each node we need some kind of hardware and software so that the nodes will have grooming capability. The ADM (Add/Drop Multiplexers), Optical ADM (OADM), Optical Cross Connects (OXC), etc which are required for grooming. Due to the requirement of these devices and additional software's at each node the cost of network goes on increasing. The sparse grooming is a technique in which few nodes will be equipped with grooming capabilities and few nodes without grooming capabilities. Due

to reduction in the number of nodes of grooming capability cost of network will be reduced. But reducing the cost is not the only objective of the research community. The main target is to get the good or acceptable performance in minimum cost so that affordable solutions and services will be provided to the end user.

In the backbone network when the grooming is provided it may result into the cost effective best performance network. But sometimes due to some reasons different types of faults may occur and the performance degrades. Faults may be due to failure of nodes or failure of links. Therefore, it is very important to provide survivability to sparse grooming in optical WDM mesh network. Survivability is provided to the network so that it becomes fault tolerant system.

Main contribution of this paper is fault tolerant system for sparse grooming in optical WDM mesh network using combiner queuing. Work focuses in this paper is on node failure in the optical WDM mesh network.

Rest of the paper is organized in sections and it follows as in section II we discuss literature survey, section III focused on proposed methodology, section IV deals with implementation architecture and performance of proposed methodology. Conclusions are given in section V and section VI focuses on future scope.

## II. LITERATURE REVIEW

Backbone network has moved from ring to mesh topology. Traffic grooming problem is addressed in both ring and mesh topology. The current focus of traffic grooming is on mesh topology [1-8]. Two approaches of traffic grooming are static and dynamic which are addresses by many papers [9-13]. Heuristics for Sparse Traffic Grooming in Dynamic WDM Optical Mesh Networks is proposed by S. R. Shinde et. al. [14] heuristic algorithms are proposed for G-node selection. On the selected G-node grooming will be performed which results in efficient solution. The focus was also the load balancing on the routes in the optical network. Where as in [15] Modified Multi-objective Meta-heuristics is proposed for Sparse Traffic Grooming in Optical WDM Mesh Networks. Any of these works is not addressing the problem of survivability over the WDM optical Mesh network.

Survivability in optical network is addressed in [16] and the focus of the work is classification of disaster survivability in optical networks. Authors have classified the survivability on the basis of present work and they have focused on survey of that classification. They also classified disasters and addressed impact of classification on communication networks. Failures may be anything of node failure, link failure, channel failure, or software failure. To recover from fault there are two solutions one is protection and other is restoration [17].

In protection method dedicated resources are required. For restoring the services affected by failure in restoration the emergency capacity available within the network is utilized. Protection method is either path protection or link protection. In path protection alternate route is reserved in advance for the source and destination connection when a connection is set up. The path protection is again either dedicated or shared path protection. In dedicated path the alternate route is only for the

one route. And in shared path protection the alternate route is shared as a backup route for more than one path as failure may not occur continuously.

As in the case of path protection link protection is another situation. In link protection the link is protected as alternate link to one where connection is established. This link protection is again dedicated and shared link similar to path protection. In restoration method there are also path restoration and link restoration. In path restoration after failure the system has to find out alternate path and for that the path will be searched at run time. This is not good method for high priority data. In link restoration as well the link will be made available by search after the failure. This is slow method and loss of data may be possible.

Focus of this paper is to provide better solution for the various problems of the current system. In the literature it is observed that focus of research is either sparse traffic grooming or survivability on WDM optical network. This work has the bidirectional focus. First of all the system is made sparse traffic grooming for Poison traffic and load balancing. Upon successful sparse traffic grooming the vulnerabilities in optical network are considered and system is provided with the better solution as a fault tolerant system over sparse traffic grooming.

## III. SURVIVABLE SPARSE TRAFFIC GROOMING USING COMBINER QUEUES

Traffic grooming is the technique in which all nodes have the capability of wavelength conversion it means all nodes are grooming node (G-node). The performance of full traffic grooming network is good but the network is not cost effective. If the cost of the network is very high, then the solutions are not affordable to the end users. So it is very important to provide cost effective best solutions to the society. One such a network is sparse traffic grooming optical WDM mesh network.

In sparse traffic grooming optical WDM mesh network only some nodes has the grooming capability. Only grooming nodes perform wavelength conversion whereas the non-grooming node does not have the capability of wavelength conversion. The nodes which have grooming capability are highly important nodes in the sparse traffic grooming. So these G-nodes are placed in such a way that less number of grooming nodes must give acceptable performance which is nearly equal to full grooming. G-nodes are selected based on various factors such as nodal degree, maximum traffic node, any random m nodes, etc. But any of these methods does not give the best solutions we are using the heuristic in [14] by applying combiner Queue to the nodes.

At each node a combiner queue will be applied and each node has its own initial weight. This weight is going to be increased based on the factors of degree and traffic at that node.

### A. Algorithm : G-Node Selection with Combine Queue (CQ)

- 1) Assign the weights to each node.
- 2) Find the Nodal degree of given Node.
- 3) Increment the weights by adding the nodal degree to each node.

- 4) Generate the Poisson traffic in the network.
- 5) Count the number of requests in each queue and add that to the weight of the nodes.
- 6) If the CQ is full, that node will be declared as G-Node.
- 7) Else the nodes with highest weights are declared as G-nodes.

Once the G-nodes are decided for the 14 node NSF net topology shown in fig. 1 then data traffic will be generated dynamically with the Poisson distribution.

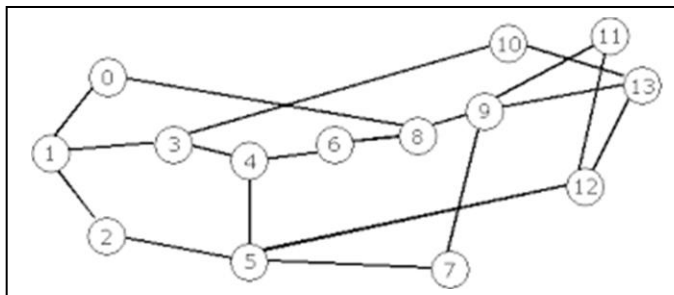


Fig. 1. 14-node NSF NET topology

System is modeled as the Markov chain model where each node is represented as one state. Rate of arrival is  $\lambda$  at each link and the service provided by the link is given by  $\mu$ . Utilization ( $\rho$ ) of the channel is given by the rate of arrival by the service rate. Due to combiner queue there are some queues at each node which are not failure say  $i$  and there are  $j$  queues where ongoing calls are there. Therefore  $\rho = \lambda / m\mu$  and the number of requests which are blocked will be given by the call blocking probability  $P[\text{block}]$  is

$$P[\text{block}] = \begin{cases} p_0 \frac{m^m \rho^n}{m!} & n > m \end{cases} \quad (1)$$

Where  $p_0$  is initial probability,  $n$  is the number of calls and  $m$  is the number of channels at that node. Here all requested are arrived at one the queue from CQ. All these request needs are analyzed and appropriate channel is assigned. If the channel is not available then the request is in wait state and the CQ will calculate the time required to get the service from that queue and other queues at that node. Then the request will be re-queued in the efficient queue of the same node. If the node is grooming node then wavelength conversion take place and probability of service in time will get increased.

If the failure of the path/link occurs then the other path/link is protected for the same. Only change here which is very important in terms of full grooming and sparse grooming is that the path/link provided as an alternate to the one selected must have nearest G-node on the path or link. Shortest path routing algorithm is used to find shortest path. The wavelength assignment algorithm as given is applied to assign the appropriate wavelength.

#### B. Wavelength assignment algorithm:

- 1) Check the Request  $R_i$  in each CQ,
- 2) if  $W_i$  available and fulfill  $R_i$  with available wavelength, assign to it.

- 3) Else if node is G-node use wavelength conversion and assign the channel to that request  $R_i$
- 4) Analyze the waiting time at each Queue of CQ at that node
- 5) Divert that request to the queue where waiting time is less of the same node and check for  $W_i$
- 6) If the waiting time is too high in all queues block that call.

#### IV. IMPLEMENTATION ARCHITECTURE AND PERFORMANCE

To evaluate our proposed methodology we have taken a network with arbitrary mesh topology. We have conducted simulation experiments on 14-node NSFNET topology with OC-192. The source and destination for the request are generated exponentially with Poisson distribution. At each node we have design Combiner Queue for each physical link.

The simulation environment is architecturally decomposed into different modules. This simulation has build in Omnet++ simulator. The node design consists of Router for routing, App module for generating source destination request and Combiner Queues.

Fig. 2 shows a Node Architecture. Queues are design to handle the requests which are arriving at the node. Router routes the request according to the shortest path to the specified destination. ‘‘App’’ design is used to generate the request in exponential order where it is sent to destination by router.

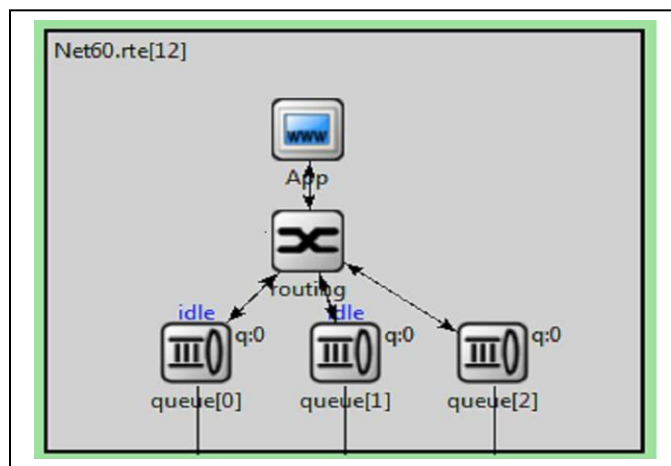


Fig. 2. Node Architecture of each node

The performance of proposed algorithm with existing few algorithms like CLPGA, MTT have been compare with parameters like Blocking probability, Network cost with Traffic Load. The proposed algorithm and MTT(Maximum Total traffic) are compared for number of sparse node selected for varying time as the traffic changes. This gives idea about how many nodes will be sparse node in the network. Based on this we have finalized number of sparse node and varying results for various parameters are achieved.

The graph for number of Sparse node selected for particular ongoing time has been plotted shown in Fig. 3

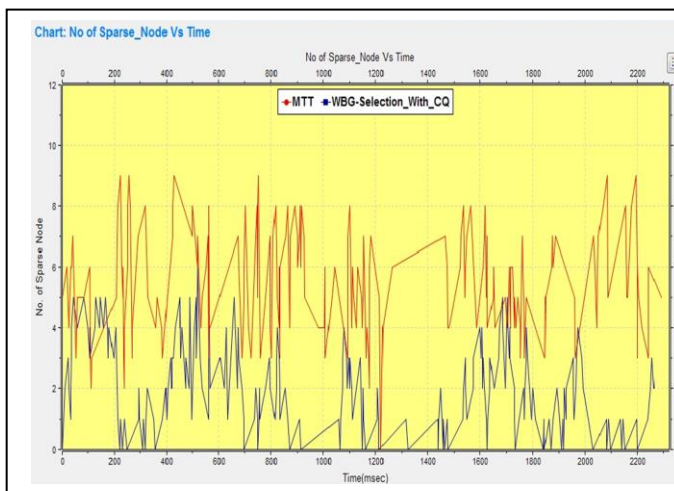


Fig. 3. No. of Sparse Node Vs Time in msec

The various algorithms are compared for the blocking probability with varying load and it is observed that the proposed method using combiner Queue gives good results as shown in Fig. 4.

Here BBORM (Bypass Based Optical Routing Algorithm) which make use of Buffering concept is compared with proposed algorithm and also the MTT has been considered over here.

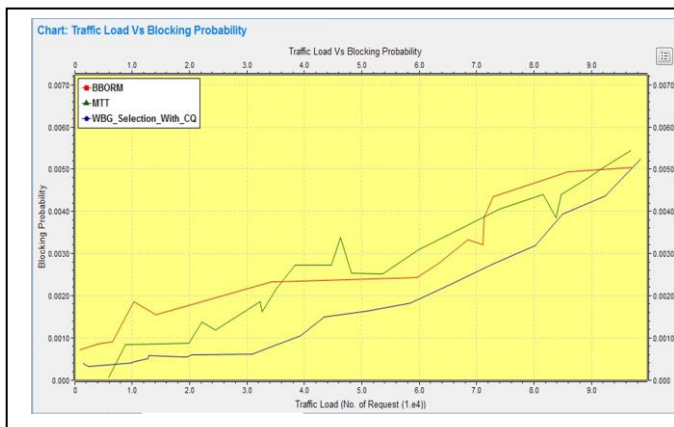


Fig. 4. Blocking probability Vs Traffic Load

Main goal of the proposed method is to reduce the network cost, and provide survivability as a fault tolerant system is achieved and is shown in fig.5.

We have seen as the load increases in the network the performance of the network degrades but in the proposed methodology the performance degradation is negligible compare to some existing methods to which we compared our results which is shown in fig.6.

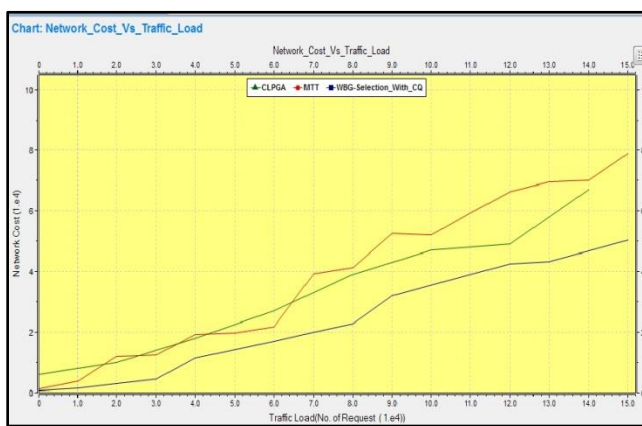


Fig. 5. Network cost Vs Traffic load

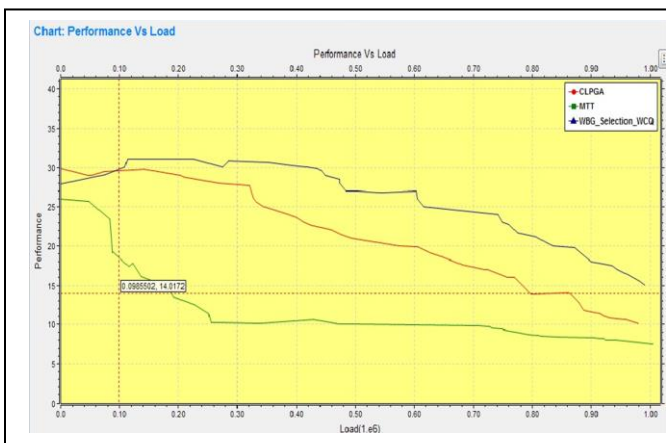


Fig. 6. Performance Vs Load

We also tried by varying the number of sparse node i.e. G-node and with varying load and we observed that blocking probability in the proposed method in any situation for any number of G-nodes is less than the available methods as shown in fig. 7.

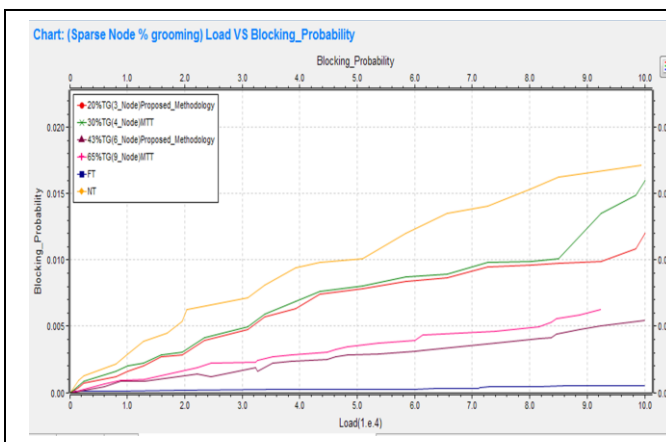


Fig. 7. Blocking Probability for Varying Sparse Node Vs Load



## V. CONCLUSIONS

Experimental work carried out using OMNet++ for fault tolerant system for sparse traffic grooming is modeled using combiner queue gives better performance by reducing the cost of network. The system is fully robust for the vulnerabilities at node failure. For fulfilling the demands of several customers over optical network Bandwidth distribution is one of the most promising ways to improve the network utilization.

The proposed methodology minimizes the total cost of traffic grooming and wavelength conversion equipment used in WDM networks with sparse traffic grooming and wavelength conversion resource without hindering the network blocking performance. We have compared our methodology with existing methodologies and from results it is concluded that our methodology is cost effective and also have less blocking probability. The GRWA problem is solved in most efficient manner in our methodology.

From results it is concluded that proposed methodology is more cost efficient and also has less blocking probability which is our objectives. Also, Load balancing over the network has been achieved by our combiner queuing model which is our one of the objective.

Our wavelength assignment algorithm strives to avoid wavelength conversion and wavelength bandwidth fragmentation by using queuing model where it is sent over respective paths by maintaining wavelength continuity constraint. The strength of the methodology lies from their applicability, minimizing the total cost of traffic grooming and wavelength conversion equipment used in WDM networks without hindering the network blocking performance.

## VI. FUTURE SCOPE

Energy consumption for traffic grooming is the future area of research which needs to be focus in order to minimize the energy requirement of network. This is also indirectly talks about the cost of the network.

Also, the power consumption for the grooming resources and minimizing the complexity of resource allocation is one of major area of research which is coming in Optical traffic grooming.

## REFERENCES

- [1] K. Zhu and B. Mukherjee, Traffic grooming in an optical WDM mesh network *IEEE J. Selected Areas Commun.* 20 (1) (2002) 122-133.
- [2] K. Zhu and B. Mukherjee, A review of traffic grooming in WDM optical networks: Architectures and challenges, *SPIE Opt. Networks Mag.* 4 (2) (2003) 55-64.
- [3] E. Modiano and P. J. Lin, Traffic grooming in WDM networks, *IEEE Commun. Mag.* 39 (2001) 124-129.
- [4] J. Q. Hu and B. Leida, Traffic grooming, routing and wavelength assignment in optical WDM mesh networks, *Proc. of IEEE INFOCOM* (2004) 495-501.
- [5] R. Dutta, G. N. Rouskas, On optimal traffic grooming in WDM Rings, *IEEE J. Selected Areas Commun.* 20 (1) (2002) 110-121.
- [6] L. Chiu and E. H. Modiano, Traffic grooming in algorithms for reducing electronic multiplexing costs in WDM ring networks, *IEEE/OSA J. Lightw. Technol.* 18 (1) (2000) 2-12.
- [7] J. Wang, V. R. Vemuri, W. Cho, and B. Mukherjee, Improved approaches for cost effective traffic grooming in WDM ring networks: ILP formulations and single-hop and multi-hop connections, *IEEE/OSA J. Lightw. Technol.* 19 (11) (2001) 1654-1653.
- [8] X. Zhang and C. Qiao, An effective and comprehensive approach for traffic grooming and wavelength assignment in SONET/WDM rings, *IEEE/ACM Trans Networking* 8 (5) (2000) 608-617.
- [9] C. Xin, Blocking analysis of dynamic traffic grooming in mesh WDM optical networks, *IEEE/ACM Trans Networking* 15 (3) (2007) 721-733.
- [10] C. Xin, Dynamic traffic grooming in optical networks with wavelength conversion, *IEEE J. Selected Areas Commun.* 25 (9) (2007) 50-57.
- [11] O. Awwad, A. I. Al-Fuqaha, A. Rayes, Traffic grooming, routing and wavelength assignment in WDM transport networks with sparse grooming resources, *Comput. Commun.* 30 (2007) 3508-3524.
- [12] C. Xin, B. Wang, X. Cao, J. Li, Logical topology design for dynamic traffic grooming in WDM mesh networks, *J. Lightw. Technol.* 24 (2006) 2267-2275.
- [13] K. Roy and M. K. Naskar, Simulation-based SONET ADM optimization approach for dynamic traffic grooming in WDM optical networks, *Photon Netw. Commun.* 16 (2008) 11-21.
- [14] S. R. Shinde, S. H. Patil, Heuristics for Sparse Traffic Grooming in Dynamic WDM Optical Mesh Networks, *IEEE International Conference on Computing Communication Control and Automation (ICCUBEA)*, (2015) page- 159 - 163
- [15] Sandip Shinde, Dr. S.H. Patil , Manasi Gulwani, Modified Multiobjective Metaheuristics for Sparse Traffic Grooming in Optical WDM Mesh Networks, *Elsevier Procedia Computer Science* 57 ( 2015 ) 980 – 987.
- [16] M. Farhan Habiba, Massimo Tornatore, Ferhat Dikbiyik, Biswanath Mukherjee, Disaster survivability in optical communication networks, *Computer Communications*, Volume 36, Issue 6,( 2013), Pages 630–644.
- [17] B. Mukherjee, “Optical WDM Networks”, Springer Publication , 2006.

# The ECG Signal Compression Using an Efficient Algorithm Based on the DWT

Oussama El B'charri, Rachid Latif, Wissam Jenkal, Abdenbi Abenaou  
Laboratory of Systems Engineering and Information Technology (LiSTi)  
National School of Applied Sciences, Ibn Zohr University  
Agadir, Morocco

**Abstract**—The storage capacity of the ECG records presents an important issue in the medical practices. These data could contain hours of recording, which needs a large space for storage to save these records. The compression of the ECG signal is widely used to deal with this issue. The problem with this process is the possibility of losing some important features of the ECG signal. This loss could influence negatively the analyzing of the heart condition. In this paper, we shall propose an efficient method of the ECG signal compression using the discrete wavelet transform and the run length encoding. This method is based on the decomposition of the ECG signal, the thresholding stage and the encoding of the final data. This method is tested on some of the MIT-BIH arrhythmia signals from the international database Physionet. This method shows high performances comparing to other methods recently published.

**Keywords**—ECG compression; wavelet transform; lossy compression; hard thresholding

## I. INTRODUCTION

The Electrocardiogram signal (ECG), as shown in Fig. 1, represents the electrical activity of the hearts. This signal is recorded using the electrodes placed on the surface of the patient's skin. The theoretical and the practical basis of the ECG recording were set out by Einthoven in 1901 [1-2]. The ECG signal contains different waves and segments .e.g. QRS complex, the P and T waves. These features represent the different sequences of the heart muscle activities, which permit to evaluate the cardiac sequences and the nature of disease affecting the heart. The ECG signal could be presented in different morphologies as presented in Fig. 2. These are due to the different way of electrodes placement [3-5].

The storage capacity of the ECG records presents an important issue in the medical practices as well as for the biomedical engineering. These data could contain hours of recording, which needs a large space for storage to save these records. The problem is more complex in the real-time processing and the transmission of the ECG signal where the doctors need to evaluate the heart condition using the fewest information possible [6]. The compression process of these records is widely used to deal with this issue. The problem with this process is the possibility of losing some important features of the ECG signal. This loss could influence negatively the analyzing of the heart condition [7-9].

The compression of the ECG signal presents a substantial challenge for searchers and engineers. Different research

works are proposed to deal with this task. Two kinds of compression are widely proposed, namely, the lossless methods and the lossy methods. With the lossless compression, the reconstructed signal is identical to the original. Different methods are proposed in this axe to deal with the ECG signal [10-14]. The disadvantages of the lossless methods are the compression rate, which does not highly minimize the storage space comparing to the original signal [15]. The aim of the lossy compression is to present a reconstructed signal close enough to the original, which permits to evaluate correctly the heart condition. This offers a high compression rate of the original signal comparing the lossless methods. Among lossy methods, the algorithms based on the discrete wavelet transform (DWT) [16-17] offer an important solution for the ECG compression. These algorithms provide a better localization of the different features consisting the ECG signal. This is due to the wavelet transform properties in the time-frequency localization [1].

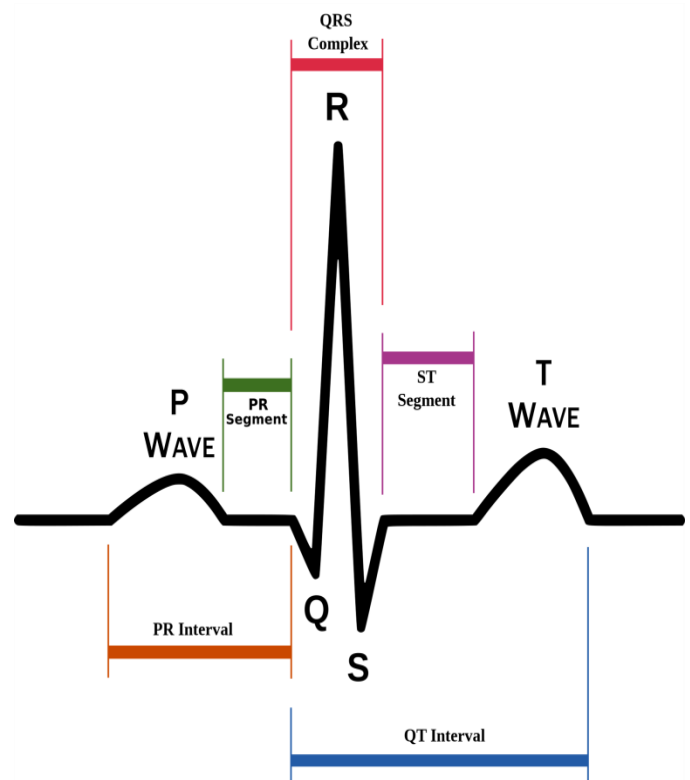


Fig. 1. Normal ECG signal with his different features

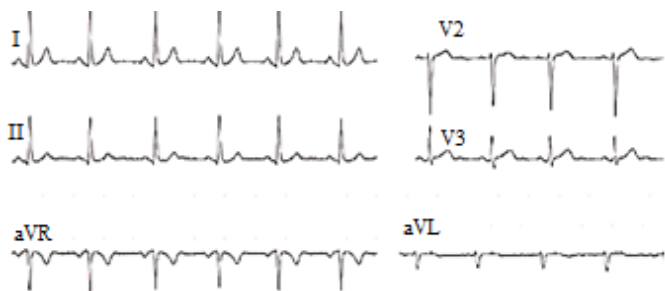


Fig. 2. Example of different morphologies of the ECG signal

This paper proposes an efficient method of the ECG signal compression using the discrete wavelet transform and the run length encoding. This method is based on the decomposition of the ECG signal, the thresholding stage and the encoding of the final data. The proposed method is tested on some of the MIT-BIH arrhythmia signals from the international database Physionet [16].

This paper is organized as follows, after the introduction; the next section presents the different steps of the proposed method from the decomposing of the ECG signal to the encoding of data. Afterward, the results and discussion section present the qualitative and the statistical results of the proposed method using some of the MIT-BIH arrhythmia signals, as well as the comparison of these results with other methods recently published. Finally, the last section concludes this paper.

## II. METHOD USED

### A. The discrete wavelet transform

The DWT is a known method widely used in the signal processing. It decomposes the signal over the different level of the high pass and low pass filters [16]. Several coefficients have been developed for large choices among different scales

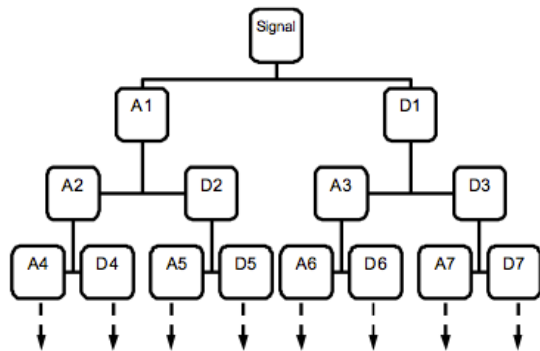


Fig. 3. The DWT decomposition

and translations in order to obtain different sorts of high pass and low pass filters. e.g. Debauchies coefficients, Symlets coefficients, Coiflets coefficients. Following are the equations of these filters for one level of the decomposition:

$$A[k] = \sum x[n] \times h(2k - n) \quad (1)$$

$$D[k] = \sum x[n] \times g(2k - n) \quad (2)$$

Where  $D[k]$  is the output of the high pass filter (detail),  $A[k]$  is the output of the low pass filter (approximation),  $h[n]$  is the half band of the low pass filter,  $g[n]$  is the half band of the high pass filter and  $x[n]$  is the discrete form of the original signal. The DWT decomposes the signal as illustrated in fig.3, where every approximation becomes a signal to be decomposed at the next level. The selection of wavelet function is a substantial step in the ECG signal decomposition [1], the selected function must be close enough to the analysed signal. The Symlet 7 (sym7) is the chosen function for this method; this function shows the best results comparing to others in the proposed method.

### B. ECG signal decomposition

This paper proposes the use of the MIT-BIH Arrhythmia signals of the international database Physionet. These signals are sampled at 360Hz, where the maximum frequency range of the real signals component is 180Hz [1]. The values of these signals are higher and positive, where the baseline of these signals is 1024 and the real values are multiplied by 200. This allows an easy conservation of the data for storage and transmission. In this paper, we propose the conservation of the MIT-BIH values. In the evaluation of the results, these values are returned to their original form, where the baseline is lower and equal to zero. The DWT decomposes the signal as illustrated in Fig. 4, where C is the vector of the decomposition, as shown in Fig. 5. This vector contains different details and the last approximation of the desired resolution. This paper proposes to decompose the signal at four resolutions, which allows observing the conservation status of the ECG features e.g. the QRS complex.

### C. The hard thresholding step

The hard thresholding permits to discard the data lower than the threshold coefficient as follows, where  $\varphi$  is the threshold coefficient and  $C[n]$  is the result vector of the wavelet decomposition:

If  $|C[n]| < \varphi$ , then  $C[n]=0$ ;  
Else  $C[n]$  is free of thresholding;

The thresholding step allows reducing the noises presented in the second approximation and set one similar value to these noises. This paper proposes an adaptive threshold coefficient as follows:

$$\varphi = \alpha \times \max_{n_1 \rightarrow n_2} |C[n]| \quad (3)$$

Where  $\alpha$  is the parameter of thresholding,  $n_1$  to  $n_2$  presents the length of the details. The proposed coefficient allows a simple thresholding of the C vector, where the  $\alpha$  parameter is the principal factor of the thresholding as shown in table 1, where the CR is the compression ratio, and the PRD is the Percentage Root mean.

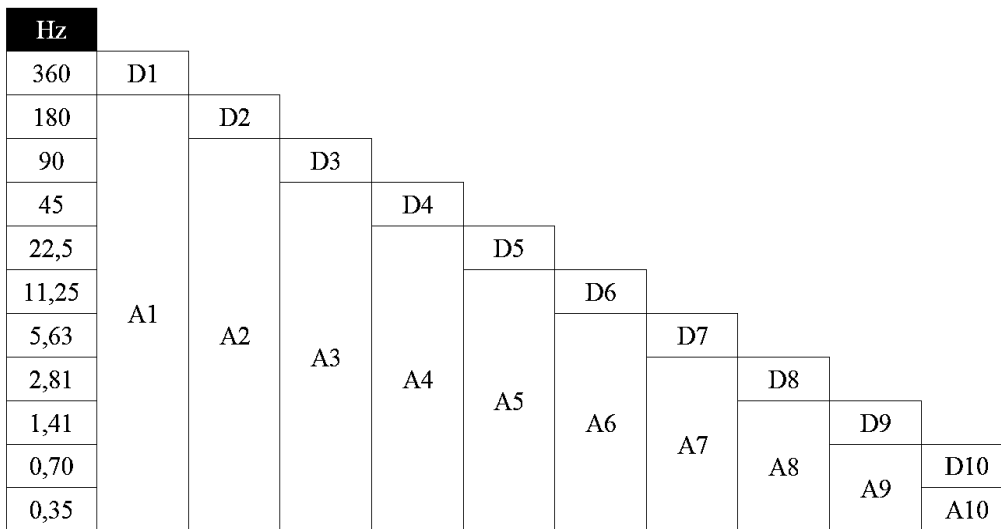


Fig. 4. MIT-BIH signals decomposition using the DWT

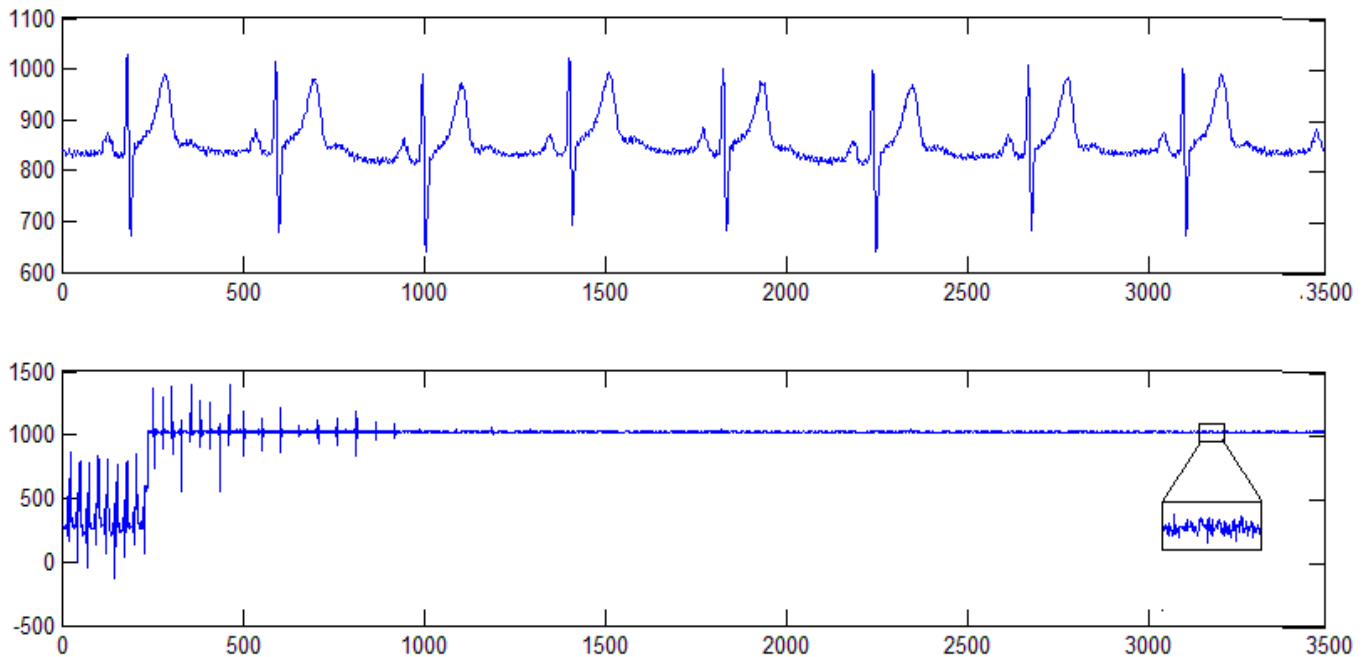


Fig. 5. MIT-BIH signals decomposition using the DWT. (a) The original signal; (b) C vector

TABLE I. THE INFLUENCE OF THE THRESHOLD PARAMETER ON THE COMPRESSION OF THE ECG SIGNAL

$\alpha$ (%)	CR	PRD
30	11,35	3,78
20	10,77	3,14
10	9,74	2,2
5	8,40	1,69

#### D. Run length encoding (RLE)

The RLE permits to minimize the length of a vector or matrix where these data show a similarity. The aim of this technique is to count the redundancy of samples .e.g. considering this vector: a,a,a,b,b,c,d,d,d the RLE data vector is: a,b,c,d and the RLE redundancy vector is: 3,2,1,3.

This paper proposes a modified RLE algorithm according to the needs of the compression process. As a result of the thresholding step, the only sample which presents a zero is concerned, where the zero presents the baseline of the signals. As indexed in the MIT-BIH database, the baseline of the ECG records is 1024. The proposed algorithm follows this property. Each presence of the sought sample is replaced by a new sample of "100". The next sample of the minimized vector includes the number the run length of the sought sample. Fig. 6 present an example of the application of the modified RLE algorithm. The proposed algorithm permits to have as a result only one vector, which his length is highly minimized compared to the original MIT-BIH signals.

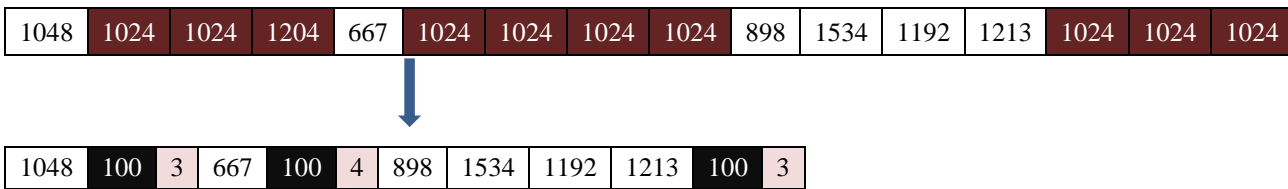


Fig. 6. The modified RLE results (a) The C vector; (b) The RLE vector

### E. The minimization of the bits number

The aim of this process is to minimize the number of data presented in the compressed vector. The MIT-BIH database contains for each signal of 10 second 3600 samples. These signals are coded in 11/10 bits. The proposed minimization process allows reducing the number of bits to 9 bits/sample or lesser. The process is simple, each sample of the RLE vector higher than 100 is divided by 4 and adding to this value the RLE indicator of 100, where only the integer results are conserved. The difference between the original RLE vector and the minimized one is very low. Other data lesser than 100 will not be minimized because these data contain the RLE parameters, which are the RLE indicator and the run length number. The last approximation could contain some negative values. Before the divided, these data are returned to a positive form by adding a fixed number of 1024 to these samples, which is close to the maximum absolute value presented in the analysed signals. These samples are excluded from the RLE process.

### F. Decompression process

The low complexity of the proposed method permits to the decompression process to be simple as well as the compression one. The decompressed process is as follows:

- Step 1. The RLE minimized vector is returned to his original form by subtracting the data higher than the RLE indicator by 100 and multiplying the result by 4.
- Step 2. The last approximation data are returned to his original form by subtracting the additive fixed number of 1024 from his values.
- Step 3. The decoding of the RLE vector using the inverse process of the modified RLE algorithm. Following is the RLE decoding process, Where RLE\_V is the RLE vector, N is the length of the RLE\_V and Rec\_C is the reconstructed C vector:

$i=1; j=1;$

For  $j=1 \rightarrow N$

If  $RLE\_V(j)=100$

- $Rle= RLE\_V(j+1);$
- $Rec\_C(i \rightarrow (i+rle))=1024 ;$
- $J=j+2 ; i=i+rle;$

Else  $Rec\_C(i)=RLE\_V(j);$

$j=j+1 ; i=i+1;$

Step 4. The last step is the inverse wavelet transform of the reconstructed C vector, which allows the reconstruction of the ECG signal.

## III. RESULTS AND DISCUSSION

The simulation results have been drawn using MATLAB. As indexed previously, this paper proposes the use of some of the MIT-BIH Arrhythmia signals.

### A. Analysis of the qualitative results

The qualitative results allow observing the signal quality before and after compression. This permits to evaluate qualitatively the performance of the proposed method.

Fig. 7 and Fig. 8 shows the results of decompression of the MIT-BIH signals n°112 and n°121 respectively. These results have been drawn using different sorts of the  $\alpha$  parameter. As presented in these figures, the quality of the decompressed signal is related to the  $\alpha$  parameter. The best result is presented in the lowest  $\alpha$  parameter. This is due to the low loss of the data in the compression process, which influence also on the compression ratio (CR). The CR for the lowest  $\alpha$  parameter is lesser than the highest  $\alpha$  parameter. So, for the highest quality, the lowest value of the  $\alpha$  parameter is recommended. However, for high compression of the ECG signal, the highest value of the  $\alpha$  parameter is recommended. Even for the 30% of the  $\alpha$  parameter, the reconstructed signal still shows important results, which is due to the high performances of the proposed method.

### B. Analysis of the quantitative results

The quantitative results permit to evaluate statistically the performances of the proposed method using different statistical parameters presented in the related works.

To evaluate the compression performances, three parameters are widely used [16-20], namely, the Percentage Root mean Difference (PRD), the Compression Ratio CR and the Quality Score (QS). Following are the PRD, the CR and the QS equations, where  $x[n]$  is the original signal,  $\tilde{x}[n]$  is the reconstructed signal, Nbo is the number of bits in the original file and Nbr is the number of bits in the compressed file:

$$PRD = 100 \times \sqrt{\frac{\sum_{n=1}^N (x[n] - \tilde{x}[n])^2}{\sum_{n=1}^N x[n]^2}} \quad (4)$$

$$CR = \frac{Nbo}{Nbr} \quad (5)$$

$$QS = \frac{CR}{PRD} \quad (6)$$

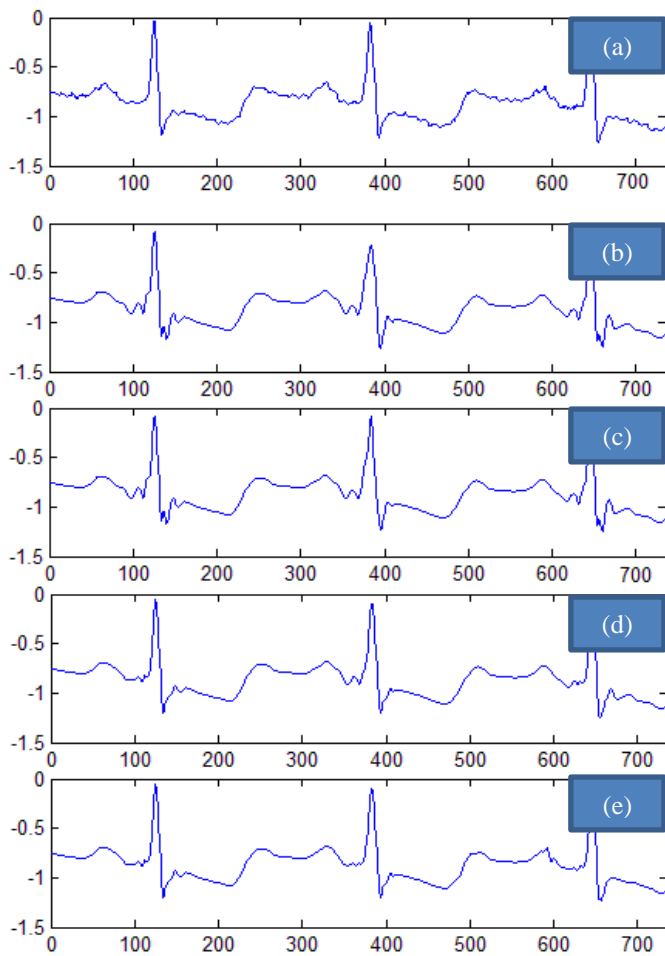


Fig. 7. Example of different morphologies of the ECG signal 112. (a) Original signal; (b) Reconstructed signal for  $\alpha=30\%$ ; (c) Reconstructed signal for  $\alpha=20\%$ ; (d) Reconstructed signal for  $\alpha=10\%$ ; (e) Reconstructed signal for  $\alpha=5\%$

The values used in the PRD evaluation are the values where the baseline is zero. This offers a real analysis of the proposed method.

Fig. 9 presents the different results of the CR parameter and the PRD parameter of different  $\alpha$  parameters in some of the MIT-BIH signals, namely, signals 112, 117 and 121. As shown in this figure, the PRD parameter varies between 5.78 and 1.62 for an  $\alpha$  parameter interval of 30% to 5%. For the same interval of  $\alpha$ , the CR parameter varies between 12.36 and 8.40.

The PRD results allow a good analysing of the ECG signal after the reconstruction. The CR results are valuable for the storage and the transmission of the ECG signal. The analysis of the statistical results allows observing the high performances of the proposed method in the compression of the ECG signal, where the proposed approach offers a simple method to deal with the storage issue of the ECG records.

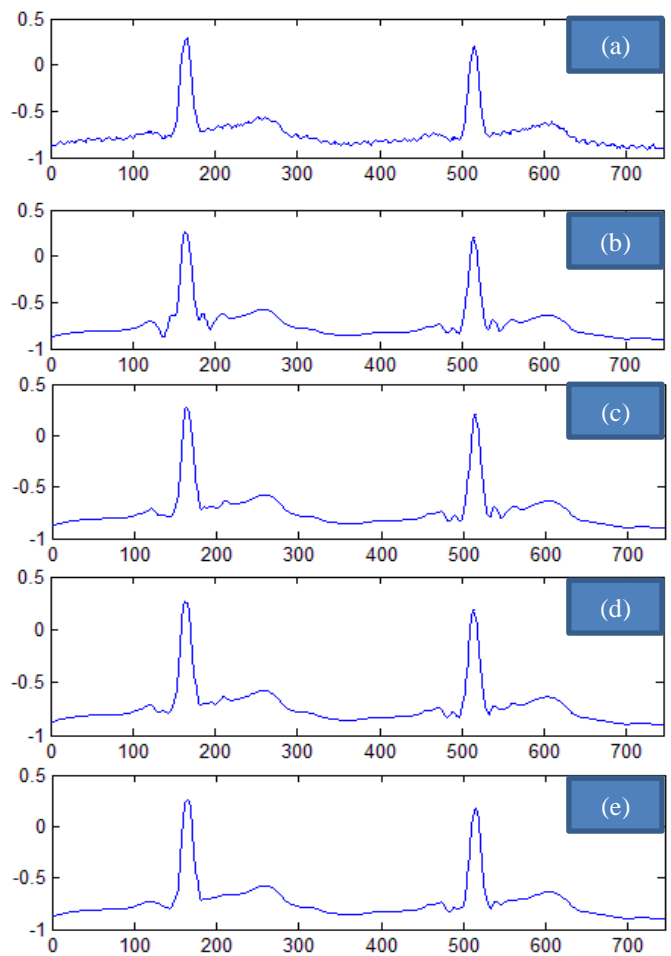


Fig. 8. Example of different morphologies of the ECG signal 121. (a) Original signal; (b) Reconstructed signal for  $\alpha=30\%$ ; (c) Reconstructed signal for  $\alpha=20\%$ ; (d) Reconstructed signal for  $\alpha=10\%$ ; (e) Reconstructed signal for  $\alpha=5\%$

### C. Performance comparison

This paper proposes the evaluation of the statistical results comparing to others techniques recently published. Table 2 presents the different results of the proposed method in the analysed ECG signals, as well as different results of other techniques, namely, R. Kumar [16], X. Wang [17-18] as reported in [16], and B. Huang [19], J. Chen [20] as reported in [19]. In bought the highest and the lowest  $\alpha$  parameters, the proposed method shows competitive results comparing to [16-20] and it shows the best quality score of 6,02.

## IV. CONCLUSION

The compression of the ECG signal is a widely used process to deal with the storage capacity of the ECG records. The problem with this process is the possibility of losing some important features of the ECG signal. This loss could influence negatively the analysing of the heart condition.





Fig. 9. CR and PRD results of the proposed method using different levels of the  $\alpha$  parameter

TABLE II. COMPARAISON RESULTS

Methods	MIT-BIH signals	CR	PRD	QS	
The proposed method	$\alpha = 30\%$	112	11,35	3,78	3,00
		117	12,08	5,78	2,09
		121	12,63	3,34	3,78
	$\alpha = 20\%$	112	10,77	3,14	3,43
		117	11,40	3,19	3,57
		121	12,13	2,48	4,89
	$\alpha = 10\%$	112	9,74	2,2	4,43
		117	10,48	2,34	4,48
		121	11,21	1,94	5,78
	$\alpha = 5\%$	112	8,40	1,69	4,97
		117	9,40	1,95	4,82
		121	9,75	1,62	6,02
R, Kumar [16]	Beta (2nd)	112 & 117	5,51	2,64	2,09
	Beta (3rd)	112 & 117	5,64	1,41	4,01
X. Wang [17]			6,30	2,55	2,47
X. Wang [18]			7,50	2,22	3,38
B. Huang [19]			10,00	2,36	4,24
J, Chen [20]			10,00	2,28	4,39

This paper proposes an efficient method of the ECG signal compression using the discrete wavelet transform and the run length encoding. This method is based on the decomposition of the ECG signal, the thresholding stage and the encoding of the final data. This method is tested on some of the MIT-BIH arrhythmia signals from the international database Physionet. As the results of the proposed approach, the PRD parameter

varies between 5.78 and 1.62 for an  $\alpha$  parameter interval of 30% to 5%. For the same interval of  $\alpha$ , the CR parameter varies between 12.36 and 8.40, which permits to have an average quality score of 4.27. In bought the highest and the lowest  $\alpha$  parameters, the proposed method shows competitive results comparing to other technique recently published.

The proposed approach shows high performances in the compression of the ECG signal with a simple method. This permits to this method to be implemented in different sorts of the software and hardware of the ECG signal analysis systems.

#### ACKNOWLEDGMENT

We gratefully acknowledge the valuable comments of the reviewers. We owe debt of gratitude to the National Centre for Scientific and Technical Research of Morocco (CNRST) for their financial support and for their supervision (grant number: 18UIZ2015).

#### REFERENCES

- [1] S. Banerjee, R. Gupta, M. Mitra, "Delineation of ECG characteristic features using multiresolution wavelet analysis method," *Measurement*, Vol. 45, Issue. 3, pp. 474-487, 2012.
- [2] W. Jenkal, R. Latif, A. Toumanari, et al., "An Efficient Method of ECG Signals Denoising Based on an Adaptive Algorithm Using Mean Filter and an Adaptive Dual Threshold Filter". *International Review on Computers and Software (IRECOS)*, Vol. 10, No. 11, pp. 1089-1095, 2015.
- [3] W. Zareba, H. Klein, *et al.* "Effectiveness of cardiac resynchronization therapy by QRS morphology in the Multicenter Automatic Defibrillator Implantation Trial-Cardiac Resynchronization Therapy (MADIT-CRT)," *Circulation*, vol. 123, no. 10, pp. 1061-1072, 2011.
- [4] X. Liu, Y. Zheng, M. W. Phyu, B. Zhao, M. Je, and X. Yuan, "Multiple functional ECG signal is processing for wearable applications of longterm cardiac monitoring," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 380-389, 2011.
- [5] W. Jenkal, R. Latif, A. Toumanari, A. Dliou, O. El B'charri, F. M. R. Maoulainine, "QRS Detection Based on an Advanced Multilevel Algorithm", *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 7, Issue. 1, 2016.
- [6] S. K. Mukhopadhyay, S. Mitra, M. Mitra, "An ECG signal compression technique using ASCII character encoding", *Measurement*, Vol. 45, No. 6, pp. 1651-1660, 2012.
- [7] D. Craven, B. McGinley, L. Kilmartin, M. Glavin, E. Jones, "Impact of compressed sensing on clinically relevant metrics for ambulatory ECG monitoring", *Electronics Letters*, Vol. 51, No. 4, pp. 323-325, 2015.
- [8] D. Moses, C. Deisy, "A Novel Lossless ECG Compression Technique for Transmission in GSM Networks", In *Proceedings of the Third International Conference on Soft Computing for Problem Solving*, Springer India, pp. 947-958, 2014.
- [9] B. Nassiri, R. Latif, A. Toumanari, S. Elouaham, F. Maoulainine, "ECG Signal De-Noiseing and Compression Using Discrete Wavelet Transform and Empirical Mode Decomposition Techniques", *International Journal on Numerical and Analytical Methods in Engineering (IRENA)*, Vol. 1, No. 5, pp. 245-252, 2013.
- [10] S. Dhar, S. K. Mukhopadhyay, S. Mitra, M. M. Baig, M. Mitra, "Noise reduction and lossless ECG encoding", *International Conference on Control, Instrumentation, Energy and Communication (CIEC)*, pp. 210-213, 2014.
- [11] O. El B'charri, R. Latif, A. Dliou, A. Abenaou, H. Jakjoud, "An enhanced method of lossless ECG data compression using ASCII character encoding", *Second World Conference on Complex Systems (WCCS)*, pp. 643-647, 2014.
- [12] S. K. Mukhopadhyay, S. Mitra, M. Mitra, "A lossless ECG data compression technique using ASCII character encoding", *Computers & Electrical Engineering*, Vol. 37, Issue. 4, pp. 486-497, 2011.
- [13] D. Venugopal, S. Mohan, S. Raja, "An efficient block based lossless compression of medical images", *Optik International Journal for Light and Electron Optics*, Vol. 127, Issue. 2, pp. 754-758, 2016.
- [14] N. Sriraam, "Correlation dimension based lossless compression of EEG signals. *Biomedical Signal Processing and Control*", Vol. 7, Issue. 4, pp. 379-388, 2012.
- [15] S. K. Mukhopadhyay, S. Mitra, M. Mitra, "A combined application of lossless and lossy compression in ECG processing and transmission via GSM-based SMS", *Journal of medical engineering & technology*, Vol. 39, No. 2, pp. 105-122, 2015.
- [16] R. Kumar, A. Kumar, R. K. Pandey, "Beta wavelet based ECG signal compression using lossless encoding with modified thresholding", *Computers & Electrical Engineering*, Vol. 39, No. 1, pp. 130-140, 2013.
- [17] X. Wang, J. Meng, "Wavelet-based hybrid ECG compression techniques", *Analog. Integr. Circ. Signal. Process.*, Vol. 59, No. 3, pp. 301-308, 2009.
- [18] X. Wang, J. Meng, "A 2-D ECG compression algorithm based on wavelet transforms and vector quantization", *Digital Signal Process*, Vol. 18, No. 2, pp. 179-188, 2008.
- [19] B. Huang, Y. Wang, J. Chen, "ECG compression using the context modeling arithmetic coding with dynamic learning vector-scalar quantization", *Biomedical Signal Processing and Control*, Vol. 8, No. 1, pp. 59-65, 2013.
- [20] J. Chen, F. Wang, Y. Zhang, X. Shi, "ECG compression using uniform scalar dead-zone quantization and conditional entropy coding". *Medical Engineering & Physics*, Vol. 30, No. 4, pp. 523-530, 2008.

# AMBA Based Advanced DMA Controller for SoC

Abdullah Aljumah and Mohammed Altaf Ahmed

Department of Computer Engineering  
College of Computer Engineering & Sciences  
Prince Sattam Bin Abdulaziz University,  
Al-kharj, Saudi Arabia

**Abstract**—this paper describes the implementation of an AMBA Based Advanced DMA Controller for SoC. It uses AMBA Specifications, where two buses AHB and APB are defined and works for processor as system bus and peripheral bus respectively. The DMA controller functions between these two buses as a bridge and allow them to work concurrently. Depending on the speed of peripherals it uses buffering mechanism. Therefore an asynchronous FIFO is used for synchronizing the speed of peripherals. The proposed DMA controller can work in SoC along with processor and achieve fast data rate. The method introduced significant volume of data transfer with very low timing characteristics. Thus it is a better choice in respect of timing and volume of data. These two issues have been resolved under this research study. The results are compared with the AMD processors, like Geode GX 466, GX 500 and GX 533, and the presence and absence of DMA controller with processor is discussed and compared. The DMAC stands to be better alternative in SoC design.

**Keywords**—FPGA; AMBA; DMA; DMA Controller; SoC; data transfer rate; FIFO

## I. INTRODUCTION

Direct Memory access (DMA) design works with processor and reduced the load of it. DMA is a logical block to access the data of peripherals and easily to understand individually. But with other blocks and processor it is difficult to understand. DMA allows accessing the data easily without the involvements of the processor, from the devices connected to computer. Therefore it makes accessing the device memory and system memory easily and allows the processor to work simultaneously on its own job while on-going operations of memory usage are carried-out by externally connected devices. Doing this DMA made system performance to boost by allowing processor to perform more other task. Many hardware entities like desk drive controllers, sound, graphics and network cards are using DMA in many systems. It plays an important role in computers to access system memory directly. Similarly it plays an important role in embedded systems. DMA becomes an important unit of System on Chip (SoC) architecture. It offers significantly fast data transfer rate between memory and external devices connected to system [1]. The DMA performance enhances while using with the bus architecture [2].

The data is transferred between system memory and peripherals first time using the old Industrial Standard Architecture (ISA) bus by The Intel's first DMA 8237 first used in IBM PC in 1981[3]. Which supports four channel and capable to transfer data by 1.6 megabytes bitrate every second.

Each single channel can address a full 64 kilo bytes part of memory and able to transfer 64 Kbytes in one programming [4]. The ISA bus and system bus were initially identical, and later separated by ISA Bridge. The IBM AT clones CPU works at higher frequency than ISA expansion therefore it is necessary to make them separate.

Another enhancement in bus architecture introduces, is in 1992 the Peripheral Component Interface (PCI) bus architecture. The concept of bus mastering introduces (first party DMA). It means a single device at a time can access the bus and known as bus master. If multiple devices have to use the same bus the arbitrary method is used [5]. Later in 1997 the packet switching concept with full duplexed mode for interfacing between system memory and multiple devices has introduces. The bus is then called PCI express (PCIe) proposed by Intel [6]. The packet switching is used and arbitrary logic replaced in the PCIe routings switches. PCIe provides background compatibility to PCI on driver levels. The PCIe increases the band width to almost double using x1 link pair for separate transmitting and receiving channels. On other side a very useful Bus architecture is going to start using by embedded products for DMA operations called advanced microcontroller bus architecture (AMBA) in System on Chip (SoC). AMBA is registered trade mark of advanced RISC machine (ARM) Ltd [7]. The first native AMBA interfaces are introduces in late 1997 with cached cores [8]. AMBA on SoC is an on-chip interconnects specification for connecting and managing various functional blocks. It supports multi-processor designs with immense numbers of controllers and peripherals. AMBA specification defines two buses named AHB and APB and is an open standard.

Currently, AMBA is extensively used in Application specific Integrated circuit (ASIC) or SoC based portable recent mobile devices for example smart phones. DMA controller and embedded processor in SoC have close relation. The performance of processor will effected drastically for not integration DMA along with it [1]. Like a human body is a system and it can perform many tasks by much subsystem. Similarly embedded system comprises multiple functions. All these functions may perform by one or more processors in any embedded system, but the question here is that, if processor is involve in doing all transfer task, how will be the performance effects. It will be busy most of the time in sharing / transferring data. For this purpose the better alternative is DMA to improve the performance and avoid the extra burden on processor.

The DMA method for data transferring along with processor in FPGA (Field Programmable Gate Array) having

many advantages. In this approach, the specific peripherals can be selected based on applications. The memory blocks, memory controllers, buses, peripherals and peripheral controller are easily added with the embedded processor system, as a result the system become progressively more impressive and useful[9]. The FPGA based method works on the principle of transmitting and receiving DMA for writing and reading respectively. It consists of two controller state machine separately, works to achieve a moderate transfer rate [10].

In the proposed method we implemented a DMA controller for embedded applications. It uses AMBA specifications and can be used in SoC based design. This DMAC is an attempt to improve data transfer characteristics. The transfer rate is achieved significantly better as compared to AMD and ARM processors used in embedded applications. It offers advanced features while keeping the very low gate count. It keeps processor in an idle state and use DMA controller to transfer volume of data. The Virtex7 FPGA is used to implement the design [11]. Two case studies have presented to compare the performance of DMAC with processor. The study examines the techniques for optimizing performance and cost in embedded system.

## II. METHODOLOGIES

### A. principal operations

Direct memory access is method in which the data will transfer between main memory and connecting peripherals and vice versa without the involvement of processor. The principle to access data from main memory by input output device independently, DMA transfer method is shown in Fig 1. DMA can perform coping (transfer) data form intra-chip in multi-core processor and memory to memory.

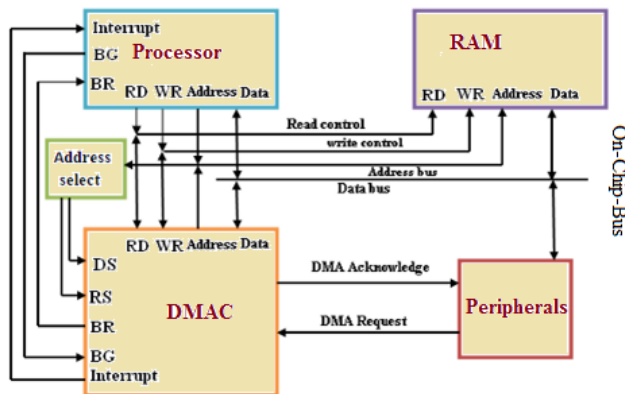


Fig. 1. DMA Transfer Methode

The logic that performs operations like addresses generation and reads writes, called DMA Controller (DMAC). The processor works in low state and configures the DMAC for data transfer, and carry on with its own task. Once processor grants the system bus, all the operations of peripherals are performed by DMAC.

Data transfer operations begins when DMAC send the bus request (BR) signal to processor to relinquish control of the

bus. In response to BR, The processor first completes the task in hand and sends the bus grant (BG) signal. After receiving BG signal, the DMAC get control on system bus and generate essential signals to perform data transfer operations. DMA controller and the memory are address by the address bus. The register selects (RS) and device select (DS) lines are activated by addressing DMA controller and are perform read or write operations at the selected memory location. The acknowledge line is set by DMA when the system starts transfer operations. Data bus transfers data between external peripherals and memory. The interrupt signal is used to inform the processor about the termination of transferring data operations. The proposed method works on the similar principle.

### B. Proposed Architecture

This research study proposes the DMA controller which works on AMBA specification. The AMBA based system usually consists of Advanced High-performance Bus (AHB) and Advanced Peripheral Bus (APB). Therefore a bridge between these two is required for transfer operations and called

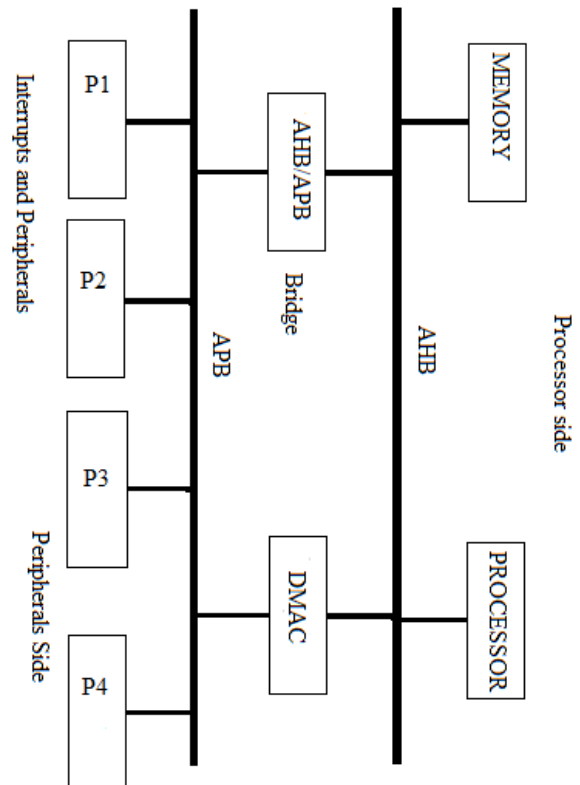


Fig. 2. AMBA System

AHB-APB bridge, this AMBA System shown in Fig 2. AMBA is open standard for 32-bit embedded processor and used in various SoCs as an on-chip system bus.

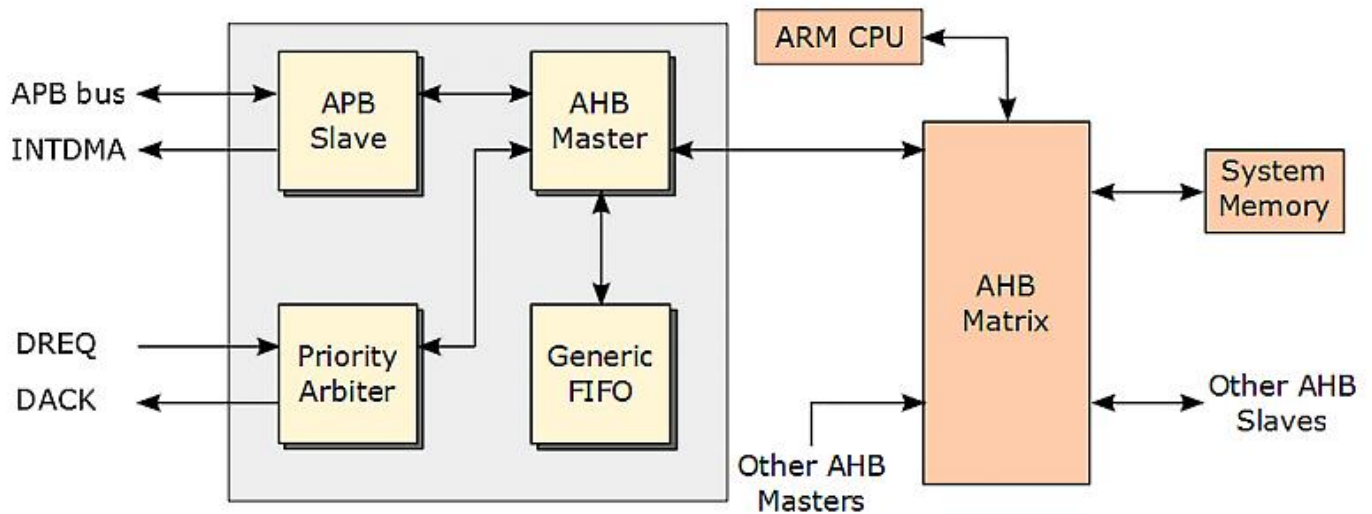


Fig. 3. AMBA Based DMA Controller for Embedded Applications

Now a day it works the de facto standard for SoC design [9]. It delivers variety of transactions like single and burst transfer, in which the single and several data packets are transferred respectively. The AMBA system usually consist of various components like memory (on chip RAM), processor, AHB/APB Bridge and peripherals like Interrupt ,UART, Timer and GPIO(General-purpose input/output).

### C. System Bus and Peripheral Bus

There are two system buses are contributes while transfer operation carried out. First one is AHB and second one ASB (Advanced System Bus). The AHB supports up to 128 bits transfer and used in high performance systems. It can support both single and multiple bus masters. It consists of master, slave, decoder and arbiter. It is available first time in AMBA 2 and later it is upgraded in AMBA 3. The simple AMBA system consists of one or more masters connected to one or more slave devices. The multiple master configuration uses arbitrator logic to choose one among several at a time. The One master multiple slaves configuration is called AHB-Lite.

The Advanced System Bus (ASB) is also a high performance bus. It also uses arbitrator for multiple masters. It is a synchronous bus, allows only one master at a time. ASB support a rich feature of pipeline where address and data transfer take place in parallel.

On the other side The peripheral bus known as APB(Advanced Peripheral Bus) is a low performance bus used to connect the peripherals to system bus of SoC. APB is interfaced with system bus (like AHB) through bridge called AHB-APB bridge for transfer operations. It allows the AHB master to address the slave on the other side of peripheral. It guaranteed about the connection but between master and slave but not ensure of correctness.

The Functional block AMBA based DMA controller for embedded applications shown in Fig 3. It consists of ARM processor, system memory, a generic first in first out (FIFO), priority arbitrator, AHB matrix and APB slave. This DMA

supports AMBA AHB-Lite. It support memory to memory, memory to peripherals and peripheral to memory transfer. This DMA Controller mainly divide in two part one side is AHB and other side is APB. DMA usually can work in one master and many slaves mode (like AHB-Lite protocol) and multiple masters and multiple slaves' mode. For this design it works as one master and multiple slave mode and can be enhance to multiple master mode in future. Therefore it is compatible with AHB-lite protocol where one master with multiple slaves' are used and the configuration is shown in Fig 4. This configuration not uses the arbiter block to select master. It contains a decoder, which decides master requests into slave triggering signals. Additionally, if there are multiple masters, an arbiter is included which decides which master gets to access the bus at any time. AHB masters can carry out burst transfers where multiple data elements are read/written from/to a slave in a particular transaction.

Multiplexer is used to confirm that only one slave from several can access the data bus at a single time. Decoder is used for selection of the slave to perform transfer operation from several. It selects the slave and set the select input of multiplexer simultaneously to select respective slave to read

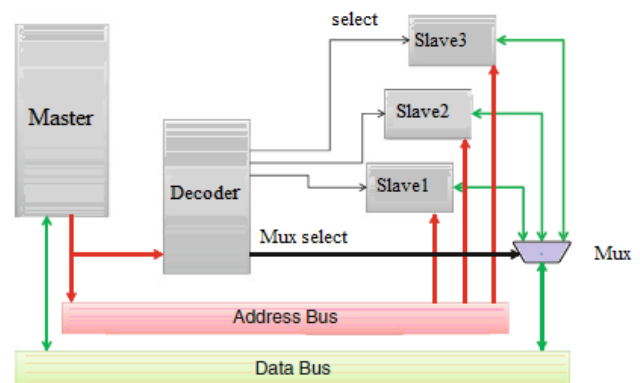


Fig. 4. One Master Multiple Slave Configuration

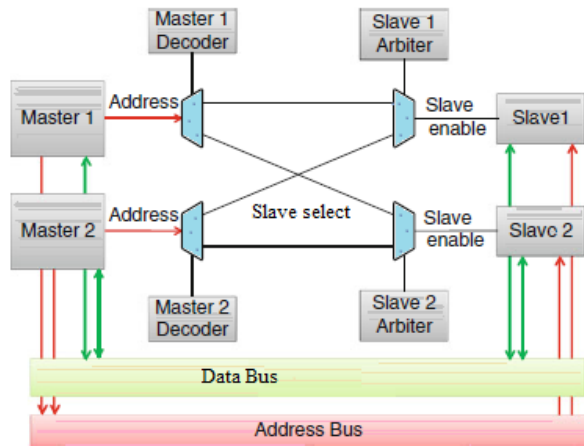


Fig. 5. Multiple Masters and Multiple Slave Configuration

from and write to data bus. DMA can also work in multiple masters and multiple slave mode, this configuration is shown in Fig 5.

This mode of operation uses arbiter logic to select one from many masters. Multiple masters can request to arbiter but arbiter grant request to only one. This method offers pipeline protocol, where address transfer, data transfer and arbitration can take place instantaneously.

DMA operation starts when the processor configures the DMA, AHB master catches the access to the bus when the arbiter grant the request of an AHB master. After receiving access the AHB master completes the data transmission between AHB and FIFO. The APB master request to achieve the control of APB, It will get access to the bus and complete the data transfer between APB and FIFO, after arbitration completes transfer with the help of the APB bridge. The APB and AHB operations are carried out independently; therefore DMAC could accomplish these two operations concurrently.

### III. RESULTS AND COMPARISON

The AMBA based Advanced DMA controller for SoC is implemented in Verilog hardware descriptive language (HDL). The simulation and synthesis are performed by Modelsim and Xilinx tool respectively. The simulation process is automated for design scheme for various test conditions under writing and reading operations, and the respective wave forms are represented in Fig 6. The transfer processes are executed upon the access to bus is approved. The hardware is synthesized, when functional verification found to be error free by modelsim. Virtex-7 target device is selected with speed grade-4. The results are achieved by successful completion of synthesis process on FPGA.

It is quite obvious from the synthesis results that the design utilizes 168 LUTs at 476MHz of maximum frequency in 2.10 nsec times. These values are tabulated in TABLE I.

The results point out to have gained maximum frequency to be 476 MHz or 476,000,000 cycles/sec, which meant that the

DMAC is able to transfer 1904 megabytes data per second [4\*476 M = 1904 megabytes/sec].

The performance results are examined by comparing the existing embedded processor of AMD. The embedded processor used in SoC such as AMD Geode GX 466 works on 333 MHz frequency, Geode GX 500 at 366MHz and Geode GX 533 on 400 MHz [12]. Therefore data transfer rate of these processors with 32 bit data width results in 1332M bytes [4\*333M = 1332 megabytes/sec], 1364M bytes [4\*366M = 1364 megabytes/sec] and 1600M bytes [4\*400M = 1600 megabytes/sec]. The comparison table is shown by TABLE II.

The data transfer rate, clearly much better in case of proposed DMA controller when compared to above three illustrations. The comparison is plotted in the graph of Fig 6.

There are few latest processor work on high frequency when compared to DMAC, obviously transfer more bytes in respect of time. However DMAC transfer more significant volume of data. Where these processors failed to maintained the same rate of transfer for large volume of data. As number of cycles increases there transfer rate decreases. The DMA functional verification is shown in Fig 7 indicating read and write operation of DMA.

Another comparison is plotted in Fig 8 for the rate of data transfer with and without DMA along with processor. Using the CPU consumes more instruction cycles to copy the data and it generally consumes more power. Instead DMA can perform the same task without processor. This way DMA is used to optimize the power and speed. Thus it is quite cleared from this comparison DMA can play a significant role in SoC for data coping. The DMAC is pleasantly incorporating significant volume of data transfer. This not only benefit through high volume data transfer, but also increased transfer rate as well, consumes low power and reducing the stress on the processor is another impact.

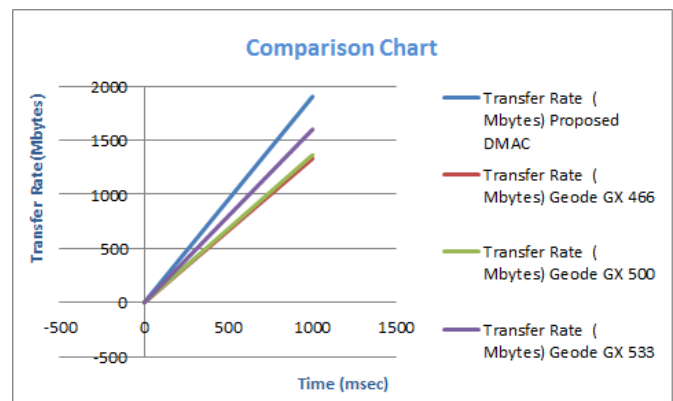


Fig. 6. Performance Comparison between DMAC and AMD Geode GX Series Processors GX 466,500 and 533



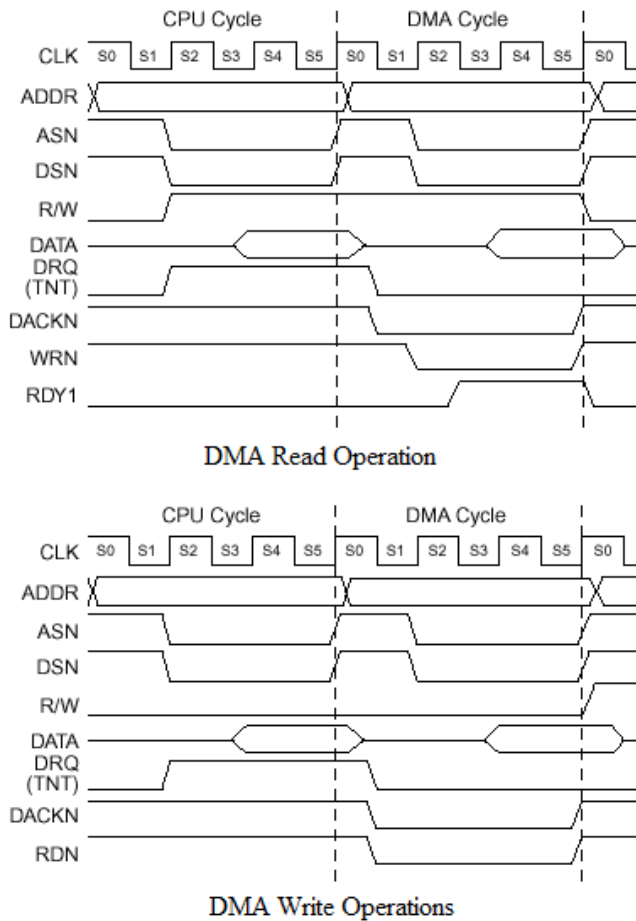


Fig. 7. DMA Read and Write Operation

IV. CONCLUSION

The AMBA based Advanced DMA controller for SoC is supposed to be a good alternative to use in SoC based embedded design. This architecture is an effort to increase data transfer characteristics. The timing and volume of data transfer are serious complications. The DMA controller has fixed these two issues. This is emphasized with three cases of comparison.

After preparing out the characteristics of these examples, the proposed DMA controller outlooks to be better alternative for high speed data transfer in innumerable application fields such as multimedia processing. Future improvements include the application of the proposed practice to various multi-processor cores where speed and power are desired.

TABLE I. SYNTHESIS RESULTS GOR DMAC

Design	Max Frequency (MHz)	Time (nsec)	LUTs	Area Utilisation	Cycles per Sec
DMAC	476	2.10	168	>1%	476 mega

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

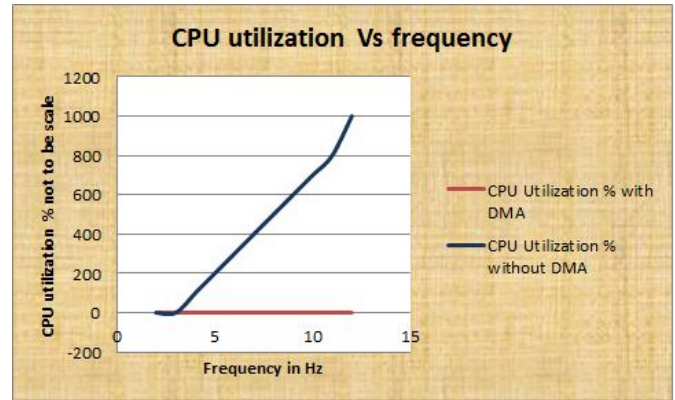


Fig. 8. Processor Utilization(instruction cycles) with and without DMA

TABLE II. TRANSFER RATE PER UNIT TIME

Transfer Time (msec)	Transfer Rate ( Mbytes)			
	Proposed DMAC	Geode GX 466	Geode GX 500	Geode GX 533
0.01	0.01904	0.01332	0.01364	0.016
0.1	0.1904	0.1332	0.1364	0.16
1	1.904	1.332	1.364	1.6
10	19.04	13.32	13.64	16
100	190.4	133.2	136.4	160
1000	1904	1332	1364	1600

ACKNOWLEDGMENT

This project was supported by Deanship of Scientific Research, Prince Sattam bin Abdul Aziz University under the research project No. 2015/01/4320.

REFERENCES

- [1] Sachin Gupta, Applications Engineer Sr, and Lakshmi Natarajan, Applications Engineer Sr, Cypress Semiconductor Corp. Published in EE Times Design (<http://www.eetimes.com>). Optimizing Embedded Applications using DMA, November 2010, pp-1-6.
- [2] Ahlem Zayati, Frédérique Biennier., Mohamed Moalla, Youakim Badr Springer Journal of Intelligent Manufacturing, "Towards lean service bus architecture for industrial integration infrastructure and pull manufacturing strategies", February 2012, Volume 23, Issue 1, pp 125-139.
- [3] Lewis, Peter H. (1988-04-24). "Introducing the First PS/2 Clones". The New York Times. Retrieved 6 January 2015.
- [4] Barry B The Intel microprocessors Brey Architecture, Programming and Interfacing. Prentice-Hall international, Inc. Fourth Edition 1997. pp-469.
- [5] Craddock, David (New Paltz, NY); Glendening, Beth A. (Poughkeepsie, NY); Gregg, Thomas A. (Highland, NY); Greiner, Dan F. (San Jose, CA) Computer Companies; "Pci Function Measurement Block Enhancements" in Patent Application Approval Process (USPTO)

- 20150261715). Investment Weekly News (Oct 10, 2015): 1040. NewsRx LLC.
- [6] Tao Jiang, Rui Hou, Jian-Bo Dong, Lin Chai, Sally A. McKee, Li-Xin Zhang and Ning-Hui sun Adapting Memory Hierarchies for Emerging Datacenter Interconnects. *Journal of Computer Science and Technology* 30(1): 97–109 Jan. 2015. DOI 10.1007/s11390-015-1507-4.
- [7] AMBA specifications, Obtained under URL: <http://www.arm.com/products/system-ip/amba-specifications.php>, ARM ltd, 2015.
- [8] David Flynn ARM, AMBA: Enabling Reusable on chip Designs. *IEEE Micro*- July-Aug 1997 pp-20-27.
- [9] J. M. Weber and M. J. Chin, Using FPGAs with Embedded Processors for Complete Hardware and Software Systems , *Lawrence Berkeley National Lab, 1 Cyclotron Road, Berkeley, CA 94720* .
- [10] Abdullah Aljumah and Mohammed Altaf Ahmed, “Design of High speed Data Transfer Direct Memory Access Controller for System on Chip based Embedded Products”.*Journal of Applied Sciences*, 15(3): Jan 2015, pp. 576-581.
- [11] Mohammed Altaf Ahmed, D. Elizabeth Rani and Syed Abdul Sattar. FPGA Based High Speed Memory BIST Controller For Embedded Applications , *Indian Journal of Science and Technology*, Vol 8(33), DOI: 10.17485/ijst/2015/v8i33/76080, December 2015 ,ISSN (Print) : 0974-6846 , (Online) : 0974-5645
- [12] White Paper AMD Geode™ GX and LX Processors Typical CPU Core Power Consumption Determination, 32077C - April 2006. (Specifications of AMD Geode).

# Role Based Multi-Agent System for E-Learning (MASeL)

Mustafa Hameed, Nadeem Akhtar, Malik Saad Missen

Department of Computer Science & IT  
The Islamia University of Bahawalpur  
Baghdad-ul-Jadeed Campus, Bahawalpur, PAKISTAN

**Abstract**—Software agents are autonomous entities that can interact intelligently with other agents as well as their environment in order to carry out a specific task. We have proposed a role-based multi-agent system for e-learning. This multi-agent system is based on Agent-Group-Role (AGR) method. As a multi-agent system is distributed, ensuring correctness is an important issue. We have formally modeled our role-based multi-agent system. The correctness properties of liveness and safety are specified as well as verified. Timed-automata based model checker UPPAAL is used for the specification as well as verification of the e-learning system. This results in a formally specified and verified model of the role-based multi-agent system.

**Keywords**—Information Management System (IMS); Multi-Agent System (MAS); Role Based Multi-Agent Systems; Agent-Group-Role (AGR); Agent-based Virtual Classroom (AVC); Intelligent Virtual Classroom (IVC); E-Learning; Information and Communication Technologies (ICTs); Formal verification; Model Checking

## I. INTRODUCTION

Use of information and communication technologies plays a vital role in a virtual classroom. E-learning systems are widely being deployed by many universities. Several new techniques like blended learning, peer-to-peer learning, and collaborative learning are being introduced [21]. Collaborative learning is an approach which builds up confidence in students and helps students to be independent in learning as well as an efficient team member to build up a knowledge base [17]. It also helps to learn from the experiences of others. This is practically possible in real classrooms but intelligent systems can also provide this as a feature in virtual classroom systems for e-learning.

Agent-based computing system developed for e-learning has concurrent processes and is distributed in nature. Along with these characteristics this type of system demands intelligence and social learning abilities. A multi-agent system is equipped with these features; and provides many fringe benefits like *autonomy*, *social ability*, *reactivity* and *pro-activeness* [3]. A role-based approach for multi-agent systems based on AALAADIN meta model [18] called AGR (Agent / Group / Role) have potential to prototype components of an organization (i.e. persons, environment) in the form of intelligent agents [1].

At the same time, the software systems are error prone [4]. And if these systems are distributed and concurrent, errors are enormous due to complexity. There are many techniques for error detection, error removal, and error reduction in software systems like software testing, inspection, or simulation [5]. Model checking is a popular technique for formal verification which formally verifies system properties to prove the correctness of the system. This enables an Intelligent Virtual Classroom with correct intended behavior.

In our paper, Section 2 elaborates state-of-the-art, agents, multi-agent systems, e-learning, correctness, formal verification, model checking, and timed-automata. In Section 3, we propose a role-based multi-agent system for an Intelligent Virtual Classroom (IVC) to be used in e-learning environment.

Our IVC is the core of *MASEL* (Multi-Agent System for E-Learning) featuring collaborative learning in an intelligent and flexible manner with Agent-Group-Role (AGR) architecture. We have elaborated on how a new skill is coined by the teacher; how students share knowledge; how students consider each other as peers; and how group leaders are selected using role-based intelligent agents. Furthermore, in Section 4, the states of the systems are defined in timed-automata and correctness of the system is formally verified with the help of model checking. Concluding notes about our proposed approach and suggestions for future work are presented in Section 5.

## II. STATE OF THE ART

### A. Agents and Multi-Agent Systems

An agent “is an encapsulated computer system that is situated in some environment” [39]. An agent-based system has autonomy, social ability, reactivity, and proactiveness properties [39][27]. Multi-agent systems may have heterogeneous or homogeneous agents. Multiple interacting agents in some environment for a common goal are called multi-agent systems, which are true implementation of agents. Agents in multi-agent systems can either have their own personal goals or an overall system goal. According to [20] multi-agent systems evolved from distributed artificial intelligence. Multi-agent systems are attractive for designing and implementing open and distributed systems because of their modularization and abstraction capabilities. Odell has noted several applications of multi-agent system [35]. ZEUS,

JADE [6], agenTool, RETSINA, JATLite, FIPA-OS, JAFMAS, Agent Building Shell, OAA, Cougaar, AgentSpace, Cybele and MADkit [24] [25] are tools for multi-agent system development [4]. Multi-agent development kit was developed at LIRMM (France). It is based on a model AALAADIN. This model shapes agents as agent, group, and role in an organization. MADKIT is used for many applications like agent-based social networks or agent-based robots [24] [25].

There are various approaches for multi-agent systems based on roles of agents [10]. Almost each approach faces the common problems of heterogeneity of languages, multiple operations, and languages and security. [18] proposes a model AALAADIN which address all these issues. AALAADIN, ROPE (Role-Oriented Programming), TRUCE (Tasks and Roles in Unified Coordination Environment), Yu and Schmid's proposal, Kendall's proposal, RoleEP (Role-based Evolutionary Programming), BRAIN (Behavioral Role for Agent Interaction), Fasli's proposal, Gaia [40], TRANS (Tractable Role-based Agent prototype for concurrent Navigation Systems), RICA (RICA-J) Role/Interaction/Communicative Action, Zhu and Zhou's proposal [10], Role based BDI framework [31] are different approaches for role-based multi-agent systems development. AALAADIN meta-model, as its advocates [18], [24], [19] claim that it addresses three main problems (heterogeneity of languages, multiple applications and architecture, security) of multi-agent systems for their design and implementation. Agent, group role is the notation used for creation of organization of multi-agent systems in AALAADIN. Community/Group/Role (CGR) as a variation of AGR is also used in MadKit [24]. MadKit is a toolkit based on java for development of organization-centered multi-agent systems based on AGR (Agent/Group/Role) architecture [24]. AGRE (Agent-Group-Role-Environment), OCMAS (Organization Centered Multi-Agent Systems), MASQ (Multi-Agent Systems based on Quadrants) are extension of AGR [19].

### B. e-Learning

The most comprehensive definition of e-learning is "the use of any kind of internet or communication service or electronic device that supports learning process" [17]. One good thing for e-learning is that now it is being widely accepted by teachers and students regardless of any particular discipline of education [29]. Modern e-learning systems use artificial intelligence to make it more efficient and productive [21]. [33] proposed a collaborative learning method based on multi-agent system. [34] studied and proposed an architecture of an intelligent tutoring system to support distance learning. Tutoring systems for distance education are not new to research and academic community. There are many early examples for these types of systems like [8] and [9]. Virtual classrooms are the core of any tutoring systems. Agent technology is being widely used to model students and teachers and handle their interaction, address their dynamic and run-time needs by their novel characteristics like intelligence and autonomy. An agent-based Virtual Classroom (AVC) represents virtual professor agent, virtual student agent, and the interfaces between them. A content agent is also present in AVC to provide relevant content and handle the changes in content [37].

### C. Correctness: Safety and liveness properties

Correctness verifies that the software behavior (i.e. functionality) is exactly according to its requirement specifications. It is a mathematical property that is absolute. Thus a program is functionally correct if it behaves according to its stated functional requirements [22]. Correctness can be accessed systematically and precisely by rigorously specifying the functional requirements [41].

The fundamental correctness properties are safety and liveness. Safety property is an invariant which asserts that something bad never happens, that an acceptable state of affairs is maintained. [32] has defined safety property as a deterministic process. ERROR conditions are like exceptions which state what is not required, as in a complex system we specify safety property by directly stating what is required. The liveness property asserts "something good happens" which describes the states of a system that an agent must bring about given certain conditions [41]. Both safety and liveness complement each other and play an important role in system verification. Progress [23] is also a type of liveness property. Reachability of state in the system under study means that, a particular state is reachable. Deadlock freeness means absence of deadlocks. This is achieved by proving safety and liveness properties. It assures that system will not stop working until in a decided terminal state. Timed automata [28] is an approach for model checking in which systems states are explicitly defined and correctness of the systems is assured by these formally verifying the safety, liveness, and reachability of these states of the system.

### D. Formal verification

Formal methods are techniques based on mathematics to design and develop software systems. Formal methods include techniques like formal specification, formal verification and automatic theorem proving [7]. Formal verification is the mathematical demonstration of the correctness of a system. Formal verification, on its mathematical foundations examines the system in accordance with the given formal specification of that system [38]. If an error is in early phases of development, it can cause big losses at the later phases. Formal verification provides ways for early detection of errors.

In [41], we have proposed a mathematical model based on UPPAAL for the design and formal verification of a multi-agent based transport system. Formal verification proves or disproves the correctness properties of the system.

### E. Model Checking

Model Checking [12][13][14][15][16][28][36] is a method for automatic and algorithmic verification of finite state concurrent systems [41]. OBDD (Ordered Binary Decision Diagrams), SAT-based translation [2], Fix-point characterization of CTL [5] and timed automata are techniques for model checking.

Model checking is for finite systems but can be scaled up for complex systems (i.e. multi-agent systems) having infinite number of states. It involves the formal verification of system properties to prove the correctness of the multi-agent system.

A complex system has a large number of independent interacting components, with non-linear aggregate activity, with concurrency between components and constant evolution [41].

Model checking due to its dynamic and automatic capabilities is more suitable for multi-agent systems; it is executed as an in-depth state space exploration that is guaranteed to terminate since the given model is finite; its algorithms are used to improve the system design; it can handle a large number of states and can be used for evolving systems [22]. There are many tools for model checking, for example, MCMAS (Model Checking Multi-Agent Systems), SMV (Symbolic Model Verifier) [11], SPIN (Simple PROMela Interpreter) [26], VDM (Vienna Development Model), LTSA (Labeled Transition System Analyzer) [32], Petri-nets and UPPAAL [30].

#### F. Timed Automata

Timed automaton models finite state real-time systems. A timed automaton is a finite-state automaton equipped with a finite set of real-value clock variables called clocks, which are used to measure the elapse of time. In [41] we have detailed timed-automata.

Timed automata are described with two elements: (1) Automata and (2) the passage of time. Control states for the system and their transition are defined in automata along with the instances of the states. Time constraints are imposed on transitions of the state with the help of clocks [5].

UPPAAL [30] is a formal tool for symbolic model checking of real-time systems developed at the University of Uppsala (Sweden) and Aalborg (Denmark). In [41], we have discussed UPPAAL as well as formally designed and verified a multi-agent based transport system.

### III. ROLE-BASED MULTI-AGENT SYSTEM FOR E-LEARNING (MASEL)

A multi-agent system based on roles for e-learning named MASEL is proposed. The core of the system is Intelligent Virtual Classroom (IVC). Features proposed in MASEL are:

- 1) *Employ collaborative learning*
- 2) *Enable real-time group discussion*
- 3) *Dynamic synchronization of learning process*
- 4) *Real-time question and answer sessions between students and teachers*
- 5) *Dynamic delivery of lessons*
- 6) *Adaptive assessment*

- 7) *Study habits of the people involved in learning process*
  - 8) *Create a sense of community by virtual teachers and students*
  - 9) *Multi-platform access*
  - 10) *Using project based learning methods*
- G. Layers of MASEL

Depending upon the functionalities and roles of individual agents, the agents can form group to form a community of agents. Groups of these intelligent agents are placed in layered architecture as shown in Figure 1.

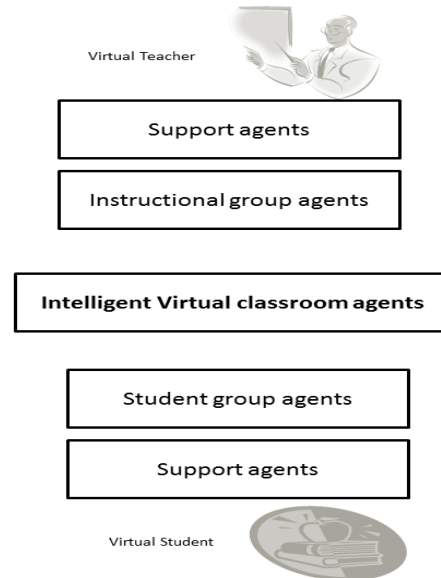


Fig. 1. Layers of MASEL

IVC agents are a group of agents which constitute the core of MASEL. Virtual students play an active or passive peer role with the real student. Virtual teachers have the ability to demonstrate a content-centered as well as student-centered role as demanded by the environment.

#### H. Roles of Agents in MASEL

Each agent in MASEL has a specific role. Based on the role of agents, they are classified in groups.

##### I. Intelligent Virtual Classroom (IVC)

The IVC consists of instructional group, student group, and support group agents. These three group of agents interact in IVC and this intelligent interaction results into a learning environment.

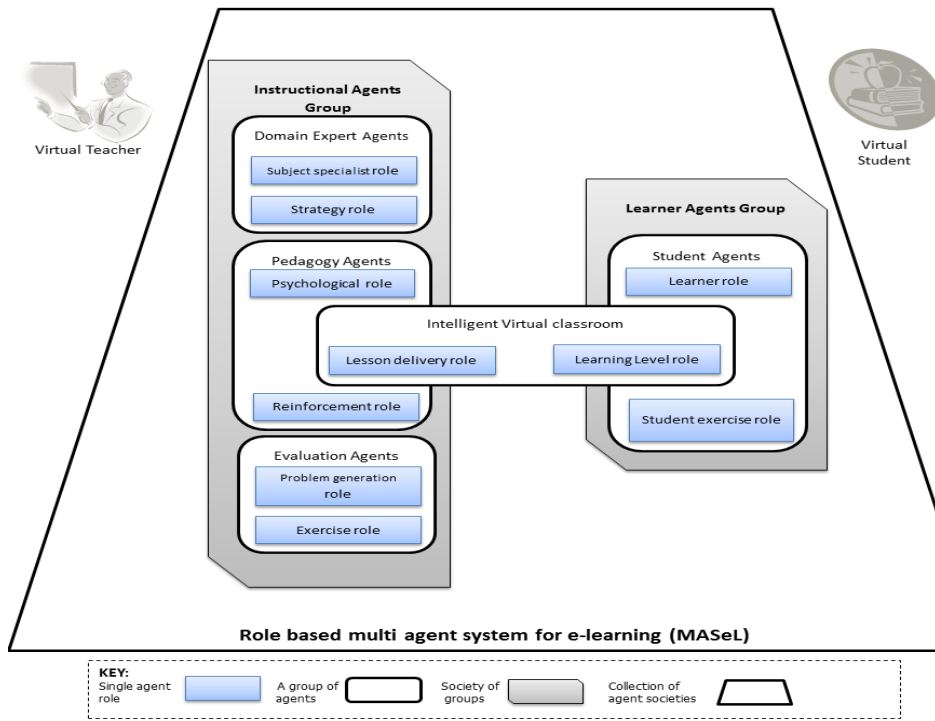


Fig. 2. MASeL: Agent roles

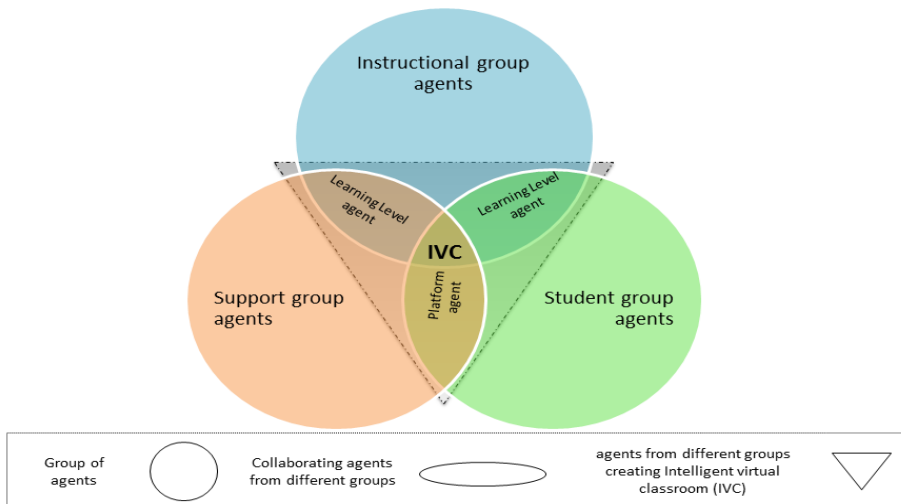


Fig. 3. MASeL: Intelligent Virtual Classroom



IV. TIMED-BASED MODEL CHECKING OF MASEL

J. Scenario-1: Active discussion session between students and teacher

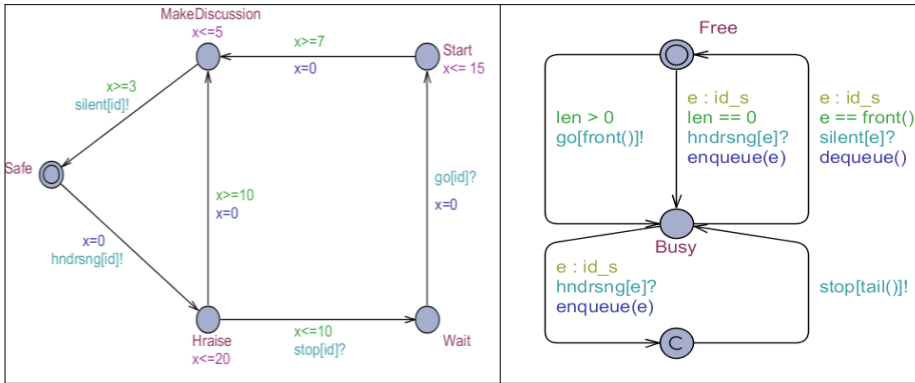


Fig. 4. UPPAAL: Timed-based automata model of active discussion session

K. Scenario-2: Student asking question in IVC

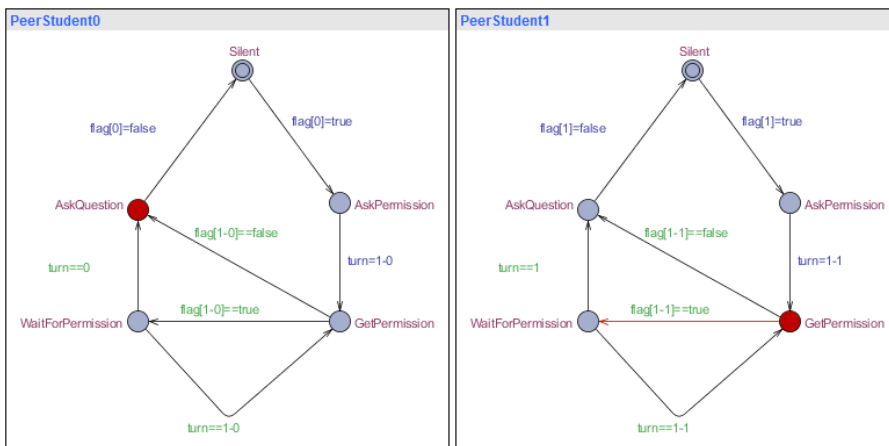


Fig. 5. UPPAAL: Timed-based automata model of student asking question in IVC

L. Scenario-3: Use of White-board by participants of IVC

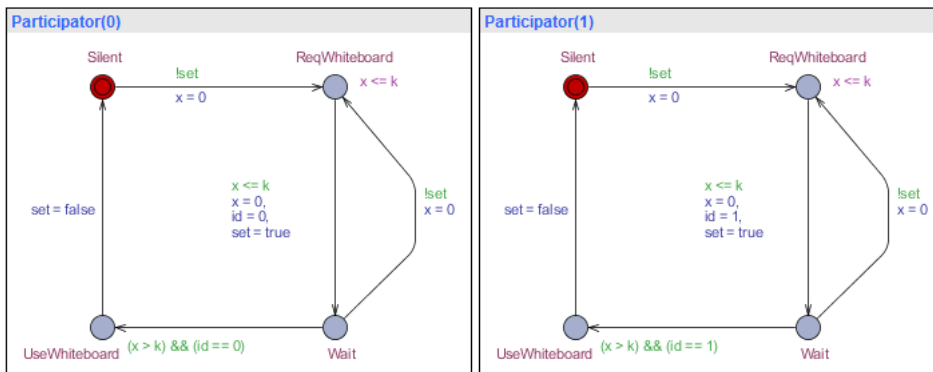


Fig. 6. UPPAAL: Timed-automata model of the use of white-board by IVC participants

M. MASEL Safety properties specified and verified in UPPAAL

Scenario	Properties satisfied by model checking	Description
White board	<pre>A[] forall (i:presenterid_t)   forall (j:presenterid_t)     Participator(i).UseWhiteboard     &amp;&amp;     Participator(j).UseWhiteboard imply i==j</pre>	At any time instance there is only one individual using the White board (Mutual Exclusion requirement)
	<pre>A[] not deadlock</pre>	The system is deadlock-free
Two students	<pre>A[( not (PeerStudent0.AskQuestion and   PeerStudent1.AskQuestion))</pre>	Mutex property
	<pre>A[] not deadlock</pre>	The system is deadlock-free
Active discussion session	<pre>E&lt;&gt; Teacher.Busy</pre>	Teacher can receive (and store in queue) messages from hand raising students
	<pre>E&lt;&gt; Student(0).Make_Discussion</pre>	Student(0) can make discussion with teacher
	<pre>E&lt;&gt; Student(0).Make_Discussion and   Student(1).Wait</pre>	When student(0) makes discussion with teacher. Student(1) waits till student(0) ends discussion
	<pre>E&lt;&gt; Student(0).Make_Discussion and   forall (i : id_s) i != 0 imply     Student(i).Wait</pre>	Student(0) can make discussion while the other students are waiting to start their own discussion
	<pre>A[] forall (i : id_s)   forall (j : id_s)   Student(i).Make_Discussion &amp;&amp;   Student(j).Make_Discussion imply i == j</pre>	At any given time instance there is never more than one student making discussion with the teacher
	<pre>A[] Teacher.list[N] == 0</pre>	There can never be N elements in the queue (thus the array will not overflow)
	<pre>Student(0).Hand_Raise --&gt;   Student(0).Make_Discussion</pre>	Whenever a Student(0) raises hand for a query, it will eventually entertained by teacher agent
	<pre>A[] not deadlock</pre>	The system is deadlock-free

V. CONCLUSION AND FUTURE WORK

Education is of fundamental importance in a developing country like Pakistan. Due to the lack of funds for educational infrastructure, less-expensive alternatives should be proposed to promote education. E-learning is one of the cheapest solutions for spreading education. And a correct e-learning system is fundamental. A model of multi-agent e-learning system that is formally specified and its correctness properties are verified is proposed in this paper. An interactive correct working model of the system is proposed. This model can be translated into a full working implementation of the system.

REFERENCES

- [1] G., Abrami, O., Barreteau, F., Cernesson. "An Agent-Group-Role based modelling framework for participative water management support". In: *Proc. of the iEMSs 3*. s.l.:s.n., pp. 497-502, 2002.
- [2] N. Amla, X. Du, A. Kuehlmann, R. Kurshan, and K. McMillan. "An analysis of SAT-based model checking techniques in an industrial environment". In *CHARME'05 Proceedings of the 13 IFIP WG 10.5 international conference on Correct Hardware Design and Verification Methods*. Springer-Verlag Berlin, 2005
- [3] C., Badica, Z., Budimac, H. D. Burkhard, M. Ivanović. "Software agents: languages, tools, platforms". *Computer Science and Information Systems/ComSIS*, 8(2), pp. 255-298, 2011.
- [4] F. M. Batista, M. G. B. Marietto, G. C. O. Barbosa, R. S. Franca, E. A. Noronha. "Multi-Agent Systems in a Computational Environment of Education: A Chatterbot Case Study". *International Journal for Infonomics (IJ) (Online)*, 3(3), pp. 285-295, issn: 17424712, 2010.

- [5] Bérard, M. Bidoit, A. Finkel, F. Laroussinie, A. Petit, L. Petrucci, Ph. Schnoebelen. *Systems and Software Verification: Model-Checking Techniques and Tools*. Springer-Verlag, 2001.
- [6] F. Bellifemine, A. Poggi, G. Rimassa, "Jade. A FIPA-compliant agent framework", In *fourth international conference on Practical Applications of intelligent Agents and Multi-agent technology (PAAM'99)* (pp. 97-108), 1999.
- [7] R. H. Bordini, M. Fisher, W. Visser, M. Wooldridge, "Verifying multi-agent programs by model checking". *Autonomous agents and multi-agent systems*, 12(2), pp. 239-256, 2006.
- [8] P. Brusilovsky, S. Elmar, and W. Gerhard, "ELM-ART: An intelligent tutoring system on World Wide Web". In: *Intelligent tutoring systems*. Berlin Heidelberg: Springer, pp. 261-269, 1996.
- [9] P. Brusilovsky, S. Ritter, E. Schwarz, "Distributed intelligent tutoring on the Web". Kobe, Japan, *8th World Conference of the AIED Society*, pp. 482-489, 1997.
- [10] Cabri, G., Leonardi, L., Ferrari, L. and Zambonelli, F., "Role-based software agent interaction models: a survey". *Knowledge Engineering Review*, 25(4), pp. 397-419, 2010.
- [11] Cimatti, E. Clarke, E. Giunchiglia, F. Giunchiglia, M. Pistore, M. Roveri, R. Sebastiani, and A. Tacchella. "NuSMV version 2: An open source tool for symbolic model checking". In *Proc. International Conference on Computer-Aided Verification (CAV 2002)*, Volume 2404 of Lecture Notes in Computer Science, pp. 359-364, Copenhagen, Denmark, 2002. Springer Berlin Heidelberg. [http://dx.doi.org/10.1007/3-540-45657-0\\_29](http://dx.doi.org/10.1007/3-540-45657-0_29)
- [12] E. M. Clarke, E. A. Emerson, "Design and synthesis of synchronization skeletons for branching time temporal logic". In *Logics of Programs* (pp. 52-71). Yorktown Heights, New York: LNCS volume-131, May 1981.
- [13] E. Clarke, O. Grumberg, D. Peled, D., *Model Checking*. MIT press, 1999.
- [14] E. M. Clarke, E. A. Emerson, A. P. Sistla, "Automatic verification of finite state concurrent systems using temporal logic specifications". *ACM transactions Prog. Lang. Syst.*, 8(2): 244-263, 1986.
- [15] E. M. Clarke, O. Grumberg, D. E. Long, "Model checking and abstraction". *ACM Transactions Prog. Lang. Syst.*, 16(5):1512-1542, 1994.
- [16] E. M. Clarke, O. Grumberg, S. Jha, Y. Lu, H. Veith, "Counter example-guided abstraction refinement for symbolic Model Checking". *Journal ACM*, 50(5): 752-794, 2003.
- [17] G. Conole, M. d. Laat, T. Dillon, J. Darby, "Disruptive technologies, pedagogical innovation: What's new? Findings from an in-depth study of students' use and perception of technology". *Computers & Education*, 50(2), p. 511-524, 2008.
- [18] J. Ferber, O. Gutknecht, "Aalaadin: a meta-model for the analysis and design of organizations in multi-agent systems", *ICMAS 98 (International Conference on Multi-Agent Systems)*, Paris, Y. Demazeau (ed), IEEE Press, pp. 128-135, 1998.
- [19] J. Ferber, T. Stratulat, J. Tranier, "Towards an integral approach of organizations in multi-agent systems: the MASQ approach". *Multi-agent Systems: Semantics and Dynamics of Organizational Models in Virginia Dignum (Ed)*, IGI, 2009.
- [20] C. E. Georgakarakou, A. A. Economides, "Software Agent Technology: an Overview Application to Virtual Enterprises". In: *N. Protogeros, ed. Agent and Web Service Technologies in Virtual Enterprises*. New York: IGI-Global, pp. 1-24, 2008.
- [21] D. Gheorghiu, L. Stefan, A. Rusu, "E-Learning and the Process of Studying in Virtual Contexts". In: *M. I. a. L. C. Jain, ed. E-Learning Paradigms and Applications - Agent-based Approach (Studies in Computational Intelligence 528)*. Berlin Heidelberg: Springer-Verlag, pp. 65-95, 2014.
- [22] C. Ghezzi, M. Jazayeri, D. Mandrioli, *Fundamentals of Software Engineering*. 2nd edition, Prentice-Hall, 2003
- [23] D. Giannakopoulou, J. Magee, J. Kramer, "Fairness and priority in progress property analysis". *Technical report*, Department of Computing, Imperial College of Science, Technology and Medicine, 1999.
- [24] O. Gutknecht, J. Ferber, "MadKit: A generic multi-agent platform". In: *Proceedings of the fourth international conference on Autonomous agents*. s.l.:ACM, pp. 78-79, 2000.
- [25] O. Gutknecht, J. Ferber, "The madkit agent platform architecture. In: *Infrastructure for Agents, Multi-Agent Systems, and Scalable Multi-Agent*. s.l.:Springer, pp. 48-55, 2001.
- [26] G. J. Holzmann, "The model checker SPIN". *IEEE Transactions on Software Engineering*, 23(5), pp. 279-295, 1997.
- [27] N. R. Jennings, M. Wooldridge, "Agent-oriented software engineering". In: *Multi-Agent System Engineering*. s.l.:Springer, pp. 1-7, 1999.
- [28] J. P. Katoen, "Concepts, Algorithms, and Tools for Model Checking". *Lecture Notes of the course Mechanized Validation of Parallel Systems*, 1999.
- [29] P. Lam, C. McNaught, J. Lee, M. Chan, "Disciplinary difference in students' use of technology, experience in using eLearning strategies and perceptions towards eLearning". *Computers & Education*, Volume 73, pp. 111-120, 2014.
- [30] K. G. Larsen, P. Pettersson, W. Yi, "UPPAAL in a nutshell". *Software Tools for Technology Transfer*, vol-1: pp.134-153, 1997.
- [31] J. Leng, J. Li, L. C. Jain, "A role-based framework for multi-agent teaming". In: *Knowledge-Based Intelligent Information and Engineering Systems*. Springer: s.n., pp. 642-649, 2008.
- [32] J. Magee, J. Kramer, *Concurrency: State Models & Java Programs*. New York: John Wiley and Sons, 2006.
- [33] H. Mazayad, I. Tnazefi-Kerkeni, H. Basson, "A Multi-Agent System to Implement a Collaborative Learning Method". In: *ICIW 2013 : The Eighth International Conference on Internet and Web Applications and Services*. s.l.:s.n., pp. 266-271, 2013.
- [34] M. T. Mitchell, "An architecture of an intelligent tutoring system to support distance learning". *Computing and Informatics*, 26(6), pp. 565-576, 2012.
- [35] J. Odell, "Agent Technology: An Overview". [Online] Available at: [http://www.jamesodell.com/Agent\\_Technology-An\\_Overview.pdf](http://www.jamesodell.com/Agent_Technology-An_Overview.pdf) [Accessed 2 February 2014], 2010.
- [36] J. P. Quielle, J. Sifakis, "Specification and verification of concurrent systems in CESAR". *Proceedings of the 5th International Symposium on Programming*, (pp. 337-350), 1982.
- [37] V. Trajkovic, D. Danco, K. Goran, Z. Petanceska, "Web-based virtual classroom". In: *Technology of Object-Oriented Languages and Systems (TOOLS)*. s.l.:IEEE, pp. 137-146, 2000.
- [38] M. Winikoff. "Assurance of Agent Systems: What Role should Formal Verification play?" Chapter 12 (pages 353-383) in *Specification and Verification of Multi-agent Systems*. Edited by Mehdi Dastani, Koen V. Hindriks, and John-Jules Meyer. ISBN:978-1-4419-6983-5, 2010.
- [39] M. Wooldridge, "Agent-based software engineering". *IEE Proceedings-software*, 144(1), pp. 26-37, 1997.
- [40] F. Zambonelli, N. R. Jennings, M. Wooldridge, "Developing Multiagent Systems: The Gaia Methodology". *ACM Transactions on Software Engineering and Methodology*, 12(3):317-370, 2003.
- [41] N. Akhtar, M. Nauman, "Timed-Automata Based Model-Checking of a Multi-Agent System: A Case Study", *Journal of Software Engineering and Applications*, Vol. 8, pp. 43-50, 2015, DOI: 10.4236/jsea.2015.82006

#### AUTHOR PROFILE

**Dr. Nadeem Akhtar** is Assistant Professor at the Department of Computer Science & IT, The Islamia University of Bahawalpur (IUB), PAKISTAN. He has a PhD from the Laboratory VALORIA of Computer Science, University of South Brittany (UBS), FRANCE with the highest honour "Tres Honorable". He has a Masters with specialization in Information System Architecture from Institut Universitaire Professionnalis e (IUP), University of South Brittany, Vannes, FRANCE. He is the recipient of Research award in year 2014 from The Directorate of Research and Development, The Islamia University of Bahawalpur. His research areas are formal methods, formal verification, multi-agent systems, system-of-systems, bigdata and self-adaptive systems.

# Detection and Identification System of Bacteria and Bacterial Endotoxin Based on Raman Spectroscopy

Muhammad Elsayeh

Systems and Biomedical Engineering  
Faculty of Engineering- Cairo University  
Giza, Egypt

Ahmed H.Kandil

Systems and Biomedical Engineering  
Faculty of Engineering- Cairo University  
Giza, Egypt

**Abstract**—Sepsis is a global health problem that causes risk of death. In the developing world, about 60 to 80 % of death cases are caused by Sepsis. Rapid methods for detecting its causes, represent one of the major factors that may reduce Sepsis risks. Such methods can provide microbial detection and identification which is critical to determine the right treatment for the patient. Microbial and Pyrogen detection is important for quality control system to ensure the absence of pathogens and Pyrogens in the manufacturing of both medical and food products. Raman spectroscopes represent a quick and accurate identification and detection method, for bacteria and bacterial endotoxin, which this plays an important role in delivering high quality biomedical products using the power of Raman spectroscopy. It is a rapid method for chemical structure detection that can be used in identifying and classifying bacteria and bacterial endotoxin. Such a method acts as a solution for time and cost effective quality control procedures.

This work presents an automatic system based on Raman spectroscopy to detect and identify bacteria and bacterial endotoxin. It uses the frequency properties of Raman scattering through the interaction between organic materials and electromagnetic waves. The scattered intensities are measured and wave number converted into frequency, then the cepstral coefficients are extracted for both the detection and identification. The methodology depends on normalization of Fourier transformed cepstral signal to extract their classification features. Experiments' results proved effective identification and detection of bacteria and bacterial endotoxin even with concentrations as low as 0.0003 Endotoxin unit (EU)/ml and 1 Colony Forming Unit (CFU)/ml using signal processing based enhancement technique.

**Keywords**—*Rapid Microbial Detection; Rapid Pyrogen Detection; Microwave Spectroscopy; Dielectric Spectroscopy; Ultra Wide Band; Cepstral Analysis; Raman Spectroscopy*

## I. INTRODUCTION

Endotoxaemia and Sepsis are the leading cause of death in surgeries [1][2]. One of its causes is the absence of an automatic system for bacteria detection and identification that provides a rapid method for bacteria and bacterial endotoxin detection to choose the suitable clinical therapeutic procedures. Such a system may represent essential tool for a quality control system in pharmaceutical industry of intravenous injection products [3][4].

Endotoxins are Lipopolysaccharide (LPS) Pyrogens produced by negative gram bacteria outer cell walls.

There are several methods for bacteria identification and classification such as gram staining, culturing and biochemical methods. Another known method such as polymerase chain reaction used to amplify short DNA fragments (Primers) that recognize sequences of genes that encode essential molecule. In case of bacteria detection and identification, the PCR method depends on the primers of DNA sequences of bacteria genes [6]. PCR needs sophisticated equipment, several components and reagents. PCR testing takes greater than 15 minutes [6].

Another drawback is the difficulty to identify and detect mixed containment. One of the new and instrumental based rapid method is the adenosine triphosphate (ATP) bioluminescence. It depends on special sample preparation utilizes specific enzyme combination and surfactant. The enzyme breaks down microbial ATP and produce visible light which measured the microbial presence. One of main drawback is the non-microbial ATP reaction and false positive indication. It is time consuming method requiring 24-48 hours to be completed. There are many other instrumental based rapid methods for microbial detection [7]. However, they are rapid, they have drawbacks such as time consuming, lower specificity, false identification and quantification.

On the other hand, there are several endotoxins' assays such as Thiobarbituric acid –assay, Rabbit Pyrogen Test(RPT), Human blood test , Endotoxin Activity Assay and the well-known (gel clotting or photometric) Limulus Amebocyte Lysate (LAL) assay [1][8]. Different methods and instrumentation, used in detection of endotoxin, already exist such as Capillary electrophoresis, Laser Induced Fluorescence (CE-LIF), Gas chromatography –Mass Spectroscopy (GC-MS), Matrix – Assisted Laser Desorption Ionization –Time of Flight Mass spectroscopy(MALDI-TOF-MS), Ion trap mass spectroscopy and Fourier Transform Ion Cyclotron Resonance Mass Spectroscopy[8-10]. However, since using the above mentioned instrumental methods, is very artful and complicated. This due to the fact that as these methods need samples' preparations and derivatization as preparatory step for instrument to detect the LPS. As a result, these methods are time consuming which is not preferred especially in case of severe sepsis that cause death in less than an hour. Moreover, they need complex procedures to ensure accurate measurements and have expensive implementations.

The main objective of this research is to establish an alternative accurate method for rapid bacterial and bacterial

endotoxin detection and identification with low running cost. Towards the progress in rapid microbial method development, vibrational spectroscopes took a contribution in developing new automatic system based on the Surface Enhanced Raman Spectroscopy (SERS). SERS is in Raman Spectroscopy to enhance the Raman scatter intensity of biological sample which is very weak as it suffers from high fluorescence background due to presence of macromolecules which is fluorophore in nature. The technique mainly depends on plasmon resonance [11][12].

Raman Spectroscopy is a vibrational spectroscopy which characterizes the material based on vibrational frequencies resulting from inelastic scatters of laser source's polarization effect. Vibrational and rotational spectroscopy are linked to the main electromagnetic wave interaction properties which is the dielectric mechanisms or polarization effects(Electronic and atomic Polarization, Dipole and ionic relaxation)[13][14]. Since the materials containing charged particles produce secondary fields when it comes in contact with electric or magnetic fields. They result in conduction, polarization, or magnetization of the particles in that material. Polarization of the particles forces the material to act as a dielectric [14][15]. The permittivity is determined by dielectric material's ability to polarize its particles under the influence of an electric field [13][14]. The polarization effect or dielectric mechanism is linked to the material's interaction properties under study of the electromagnetic waves, especially electric field components of the wave. The vibrational mode of the molecule depends on the frequency of the applied field. The vibrational motion of the molecules induces a displacement in molecule equilibrium position, called electric displacement and can be expressed by equation (1)

$$D(\omega) = \epsilon_0 E(\omega) + P \quad (1)$$

Where,  $\omega$  is the angular frequency  $2\pi f$ ,  $E(\omega)$  is the electric field of the microwave relative to oscillation frequency,  $P$  is the Polarization density corresponding to dipole moment density expressed by equation (2)

$$P = \alpha E \quad (2)$$

So the displacement occurring is related to the applied electric field which induced by the dipole moment, where  $\alpha$  is the molecular polarizability which is a material property that depend on the material structure and bond nature. The polarizability has direct link with material dielectric properties as expressed in equation (3) and (4)

$$\alpha = \chi_e \epsilon_0 \quad (3)$$

Where  $\chi_e$  is the electric susceptibility of the material and can be expressed as equation (4)

$$\chi_e = \epsilon^*(\omega) - 1 \quad (4)$$

Where,  $\epsilon^*(\omega)$  is the complex permittivity relative to each frequency variation. According to equation (3) and (4),  $\alpha$  can be expressed as follow in equation (5)

$$\alpha = (\epsilon^*(\omega) - 1) \epsilon_0 \quad (5)$$

By substituting in equation (2) using equation (5), we can express the dipole moment in terms of complex permittivity and electric field relative to frequency variation in equation (6)

$$P(\omega) = (\epsilon^*(\omega) - 1) \epsilon_0 E(\omega) \quad (6)$$

In case of electronic and atomic polarization the induced dipole moment mechanism will be related to the exciting and radiated photons. Accordingly, dipole moment can be represented in different form as defined in equation (7)

$$P = \alpha_o E_o \cos(2\pi f_o t) + \frac{1}{2} \left( \frac{\partial \alpha}{\partial q} \right) q_o f_o [ \cos(2\pi(f_o + f_m)t) + \cos(2\pi(f_o - f_m)t) ] \quad (7)$$

The equation represent oscillating dipole which radiates photons with three different frequencies  $f_o$  (elastic scattering),  $f_o + f_m$  (inelastic scattering with shorter wave length) and  $f_o - f_m$  (inelastic scattering with longer wave length). Where  $f_o$  is the frequency from the vibrating molecules of the sample scattered is,  $f_m$  is the frequency of the vibrational mode and  $q_o$  is the vibrational amplitudes.

The polarizability should be changed with vibrational displacement  $q$ , the change of Raman intensity is proportional to the square of the rate change of the polarizability with respect to change of the displacement  $\left( \frac{\partial \alpha}{\partial q} \right)_o^2$ . In equation (5) the polarizability is related to dielectric properties of the material, so the intensity amplitude will change depending on the molecular composition of the sample.

The different dielectric mechanism can be cross linked with material properties and its interaction with the excitation source. As the scattering parameters are in the scope of this study, we will find a similarity of the analysis developed in the authors' previous work [14] however the excitation source is different. The paper is organized as follows, the materials and methods are described in section two. The results and discussion are included in section three. The conclusion of this work is presented in section four.

## II. MATERIALS AND METHODS

### A. Experiment Setup

The process steps and setup can be defined as shown in process diagram figure (1). The excitation source is based on a Raman Analyzer from Thermo Scientific. The Laser excitation source of 785 nm and output power of 250mW. The spectral range from 250 to 2875  $\text{cm}^{-1}$ . The sample is placed in contact with Raman analyzer focal lens about 18mm to focal plane. The actual measuring distance is 4.75 mm from outer surface of the sample test tube. The test tube is made of borosilicate glass of 16 mm in diameter and 125 mm in height. The test tube has wall thickness of 1.8 mm.

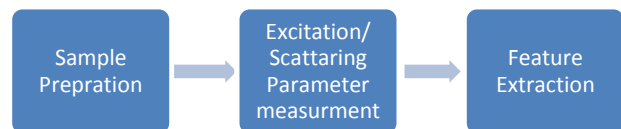


Fig. 1. Method Process diagram

**B. Samples Preparations and Measurements**

The reference samples were consisted of two water types (LAL Reagent water, Controlled distilled water free of microbial contamination and of less than 0.25 EU/ml ). The samples under test were divided into endotoxin containment and microbial containment. The serial dilutions of different concentrations were prepared by a control spike which is a lipopolysaccharide from E.coli055:B5 as certified by Charles River Endosafe® as shown in Table (1). Bacteria culturing were used to prepare different bacteria containment, we used Nutrient broth and agar slants to prepare lyophilized cultures of (Kocuria rhizophila (Micrococcus luteus) ATCC 9341, Pseudomonas aeruginosa ATCC 9027, Staphylococcus aureus ATCC 6538P, Escherichia coli ATCC 25922 and Staphylococcus haemolyticus ATCC 29970D-5.

In Raman Spectroscopy experiments, the data collected were Raman scatter signals which represented in Raman amplitude and Raman shift. Raman shift is the change in the frequency between the incident and emitted light as in equation (8)

$$\Delta\nu(\text{cm}^{-1}) = \left( \frac{1}{\lambda_{\text{incident}}} - \frac{1}{\lambda_{\text{scattered}}} \right) \times 10^{-7} \quad (8)$$

TABLE I. CONTAINMENT SAMPLES CONCENTRATIONS

Samples	Concentration EU/ml	Samples	Concentration EU/ml
1	0.0003	2	0.003
3	0.03	4	0.06
5	0.125	6	0.25
7	0.5	8	2.5
9	5	10	50
11	100	12	1000
13	10000	14	100000
15	Reference Sample		

**C. Feature Extraction Algorithm**

The algorithm depends on the nature of the output signal. The signal is defined by Raman shift and intensity. We convert the represented wave number into frequency as defined by equation (9) and apply the Cepstral transform for the signal by using equation (10).

The main goal is to find the variance between reference and contaminated sample in terms of specific signature of bacterial endotoxin and bacteria cells. Cepstral analysis algorithm is a well-known algorithm in speech recognition applications and analysis of radar signal returns applications [17]. The Cepstrum is defined as the inverse Fourier transform of the log-magnitude Fourier Spectrum [18]. Windowed liftring  $w(n)$  is applied on the Cepstrum to eliminate the highest and lowest quefrencies by using equation (11). Then, a final signature can be obtained by taking the Fourier transform of the sample signal and normalizing it by the reference signal on its Frequency domain form.

$$f = c \times \Delta\nu(\text{cm}^{-1}) \quad (9)$$

$$C_{S_{Raman}}(Q) = F^{-1} \{20 \times \log(S_{Raman}(f))\} \quad (10)$$

$$w(n) = \begin{cases} 0.5 & 1 \leq n \leq 2 \\ 1 & 2 < n < N - 20 \\ 0.5 & N - 20 \leq n \leq N \end{cases} \quad (11)$$

The detection signature can be calculated as follow

$$S_{detection} = \frac{F\{C_{S_{Raman(containment)}}(Q) * w(n)\}}{F\{C_{S_{Raman(Ref)}}(Q) * w(n)\}} \quad (12)$$

**III. RESULTS AND DISCUSSION**

Figure (2) shows the raw Raman spectrum for the lowest (0.0003 EU/ml) and highest endotoxin concentration (100,000 EU/ml) against the same Pyrogen free reference sample. It shows the variance occur to laser scattered waves due to the change in material properties as result of contamination with different concentration. The same waveform response is detected due to the strong scattering from Borosilicate glass. A weak response was detected this was shown as an absence of specific wave form for water and endotoxin at their known wave number assignments. Water cannot be detected due to device's spectral range limitation. Water Raman spectra range is from (3000 -3800  $\text{cm}^{-1}$ ). Also Lipid-A potion in LPS cannot be detected due to spectral range limitation which have strong response at 2934 $\text{cm}^{-1}$ .

Figure (3),(4) are the raw Raman spectra of the two positive gram (*staphylococcus haemolyticus* and *micrococcus luteus*) bacteria and two negative gram bacteria (*Pseudomonas aeruginosa* and *Escherichia coli*) inside distilled water. The distilled water has endotoxin concentration < 0.25 EU/ml. The results show no significant difference in wave form or intensity which show weak scatter and spectral information for bacteria samples.

The signals showed in these figures contain hidden endotoxin, bacteria Raman scattered data.

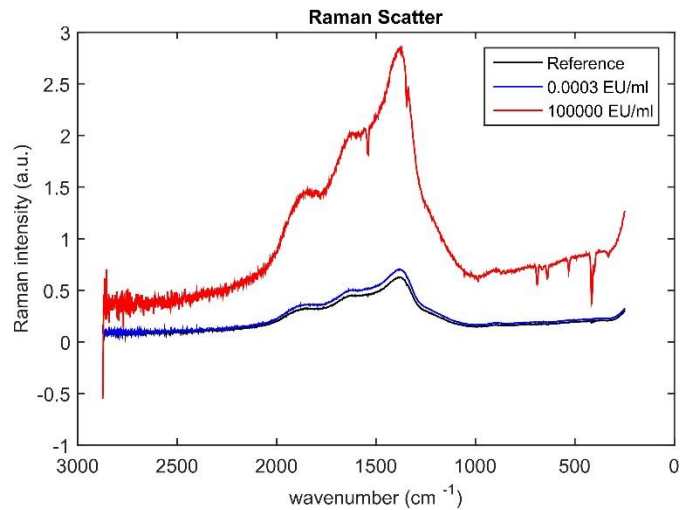


Fig. 2. Raman Spectrum for lowest and highest endotoxin concentration against reference pyrogen free distilled water



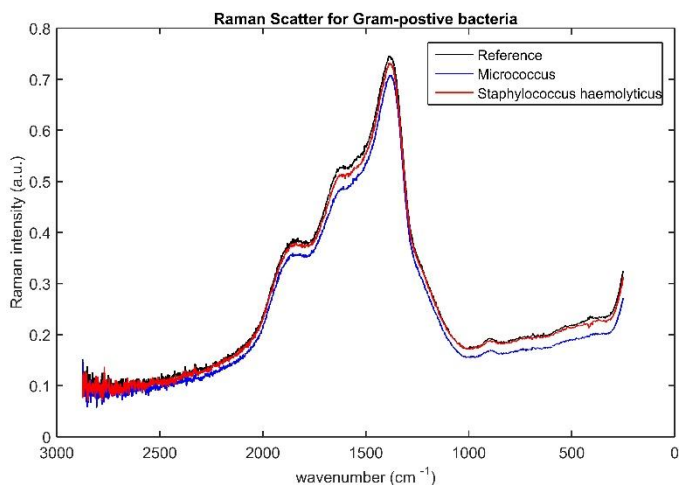


Fig. 3. Raman Spectrum for different mixture of gram Positive bacteria and distilled water with endotoxin less than 0.25 EU/ml

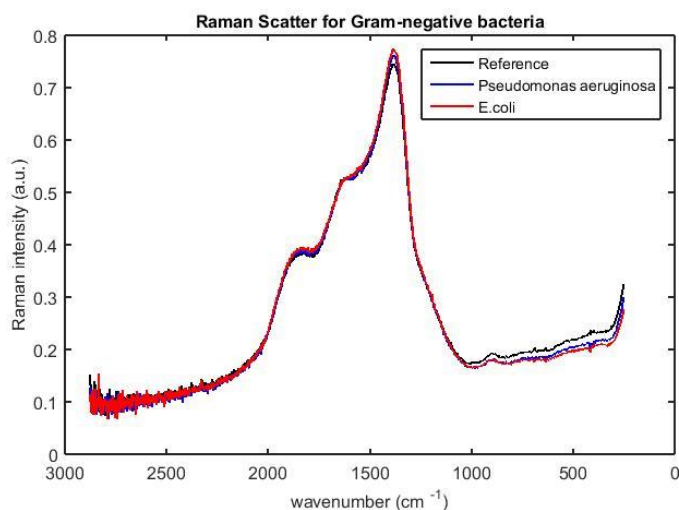


Fig. 4. Raman Spectrum for different gram negative bacteria and distilled water with endotoxin less than 0.25 EU/ml

Power spectrum representation of Raman spectrum show the specific existence of endotoxin which can be used in the bacteria classification as shown in figure (5 a). The endotoxin samples can be discriminated from the gram negative bacteria by the amplitude value and the specified frequency in the band region from 4 to 5 GHz. This can be correlated with the endotoxin material specific response in frequency bands (4 - 5 GHz) of microwave scattering parameter S11 as shown in figure (5b and 5c) [14]. This frequency band is related to actual wave number assignment known for LPS which is (1400 – 1500 cm<sup>-1</sup> equivalent to 4-4.5 GHz). These wave numbers band cover carbohydrate, lipid and protein constitutes of gram negative outer membrane and cell wall, viz, lipopolysaccharides, lipoproteins, phospholipids, proteins and porins [19]. Wave number band (1500-1600 cm<sup>-1</sup> equivalent to 4.5-5 GHz) cover carotenoid and nucleic acids [19]. The technique was challenged to determine if it can detect two mixtures of the same class of bacteria containment. In Figure (6), the output signal captured for the mixture is a result of mean value of the two main signals which represent the signature of each individual bacteria.

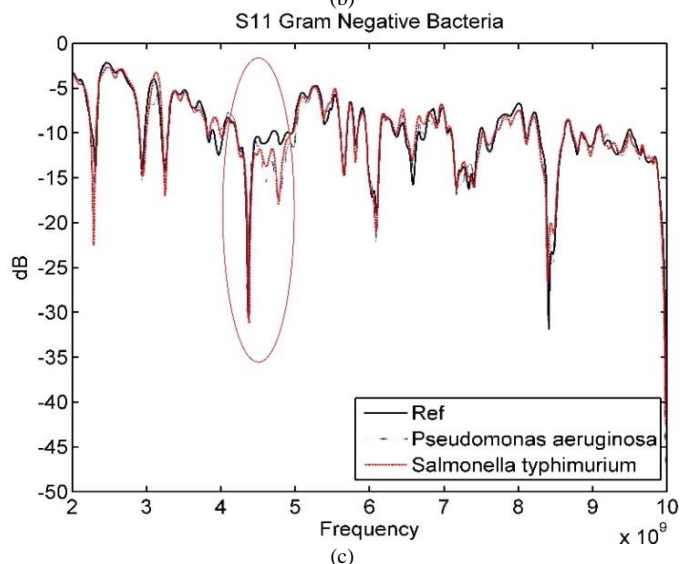
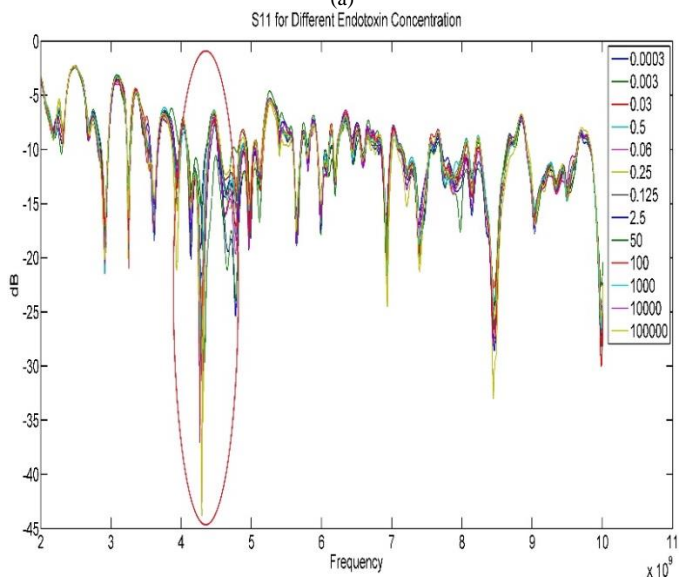
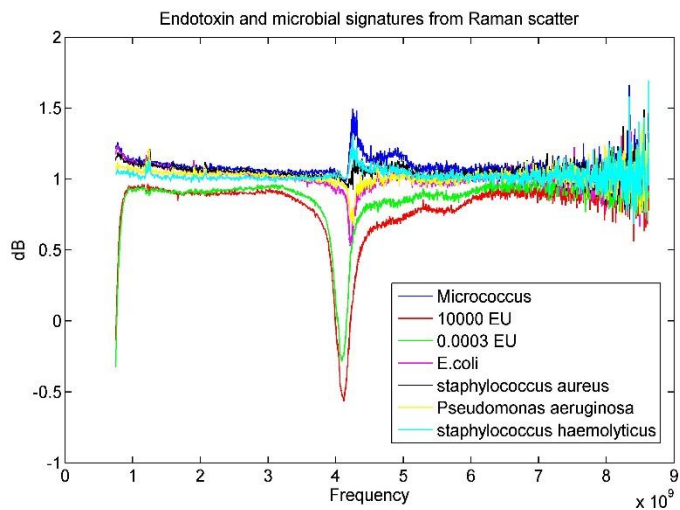


Fig. 5. a) Normalized filtered Raman power spectrum for different bacteria species and different endotoxin concentration. b) S11 signal for Different Endotoxin concentration, the circular annotation shows the specific wave

form and response from 4 to 5 GHz bandwidth [14]. c) S11 signal for different gram negative bacteria contaminant a distilled water sample with endotoxin less than 0.25 EU/ml [14]

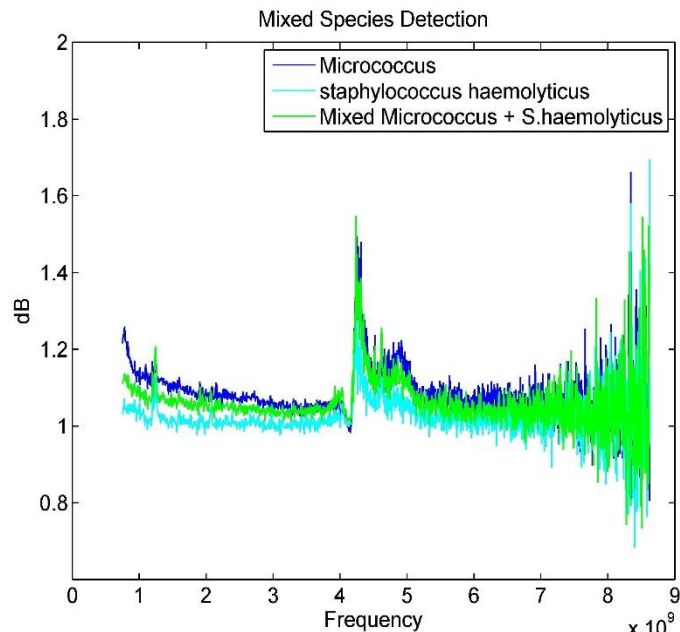


Fig. 6. Normalized filtered Raman power spectrum for different mixture of gram positive bacteria species against individual species

#### IV. CONCLUSION

A new technique for Raman enhancement using signal processing to identify and detect bacteria and bacterial endotoxin. It shows good sensitivity in LPS feature extraction and detection. It can be used as low cost, simple and rapid method for microbial and pyrogen detection applications. Also, it can be used for different bacteria identification and classification system. The method has low processing time less than 1 min.

We have proven that the frequency component of the lipopolysaccharide Raman scattered as same as the Ultra Wide Band (UWB) reflection scattered parameter measured in our previous work [14]. Although they are totally different electromagnetic excitation source but they have the same effect of standing wave caused by the interaction between the incident and scattered field that can be described in terms of reflection coefficient which, in turn, is related to the backscatter efficiency and similarities. In addition, we have offered solution for detection of bacteria mixture from different species in same medium.

#### ACKNOWLEDGMENT

We need to thank GalxoSmithKline-Giza site in Egypt, Chemistry analytical quality control Laboratory: Dr.Ahmed Ramadan and his team, Microbiology laboratory : Dr. Manal Aref and her team and Dr. Odette Boulos , head of quality for offering microbial and Pyrogen samples preparation and usage of Raman analyzer on their work place.

#### REFERENCES

- [1] B. Eisele, M. Lamy, L.G. Thijs, H.O. Keinecke, H.P. Schuster, F.R. Matthias, E. Fourier, H. Fourier, U. Delves, Antithrombin III in patients with severe sepsis, A randomized, placebo-controlled, double-blind multicenter trial plus a meta-analysis on all randomized, placebo-controlled, double-blind trials with antithrombin III in severe sepsis, *Intensive Care Medicine*, 24 (7) (1998) 663-672.
- [2] E. Esteban, R. Ferrer, L. Alsina, A. Artigas, " Immunomodulation in sepsis: the role of endotoxin removal by polymyxin B-immobilized cartridge," *Mediators Inflamm*; vol.2013:507539, (2013).
- [3] FDA, "Guidance for Industry: Pyrogen and Endotoxins Testing: Questions and Answers", <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm314718.htm>, (2012).
- [4] M. Daneshian, A. Guenther, A. Wendel, T. Hartung, S. von Aulock, In vitro pyrogen test for toxic or immunomodulatory drugs, *Journal of Immunological Methods*, 313 (2006) 169-175.
- [5] E.T. Rietschel, T. Kirikae, F.U. Schade, U. Mamat, G. Schmidt, H. Lippnow, A.J. Ulmer, U. Zähringer, U. Seydel, F. di Padova, M. Schreier, H. Brade, "Bacterial endotoxin: molecular relationships of structure to activity and function", *FASEB J.* 8, 217-225, (1994).
- [6] A.K. Järvinen, S. Laakso , P. Piiparinen, A. Aittakorpi, M. Lindfors, L. Huopaniemi, H. Piiparinen, and M. Mäki, "Rapid identification of bacterial pathogens using a PCR- and microarray-based assay" *BMC Microbiology* 2009, 9:161, <http://www.biomedcentral.com/1471-2180/9/161>, (2009).
- [7] L. Jimenez, "Molecular Diagnosis of Microbial Contamination in Cosmetic and Pharmaceutical Products: A Review", *Journal of AOAC international* Vo.84, No.3, (2001).
- [8] A. Bui, A. Kílár, Á. Dörnyei, V. Poór, K. Kovács, B. Kocsis, F. Kílár, "Carbohydrate composition of endotoxins from R-type isogenic mutants of *Shigella sonnei* studied by capillary electrophoresis and GC-MS", *Croatica Chemica Acta* ,84 (3) (2011) 393-398.
- [9] A. Bui, B. Kocsis, and F. Kílár, "Methodology to label mixed carbohydrate components by APTS", *Journal of Biochemical and Biophysical Methods*, 70 (6) (2008) 1313-1316.
- [10] A. Kílár, Á. Dörnyei, A. Bui, Z. Szabó, B. Kocsis, and F. Kílár, "Structural variability of endotoxins from R-type isogenic mutants of *Shigella sonnei*", *Journal of Mass Spectrometry*, 46 (1) (2011) 61-70.
- [11] K. Kneipp, A.S Haka, H. Kneipp, K. Badizadegan, N. Yoshizawa, C. Boone, K.E. Shafer-Peltier, J.T. Motz, R.R. Dasari, and M.S. Feld, "Surface-enhanced Raman spectroscopy in singleliving cells using gold nanoparticles," *AppliedSpectroscopy* 56 (2002) 150-154.
- [12] A. Campion and P. Kambhampati, "Surface-enhanced Raman scattering," *Chemical Society Reviews* 27, (1998) 241-250.
- [13] Jovan Mijovic and Benjamin D.Fitz, "Dielectric Spectroscopy of reactive Polymers", Department of chemical Engineering, Chemistry and Materials Science, Polytechnic University, six Metrotech Center, Brooklyn, NY 11201, Novocontrol Application Note Dielectrics 2, 1998.
- [14] M.Elsayeh, A.H.Kandil, " Ultra Wide Band Based Quantitative and Qualitative Method for Bacteria Endotoxin Detection", *Journal of Microwave Power and Electromagnetic Energy* 49 (1) (2015) pp. 11- 20.
- [15] N.N. Rao, "Elements of Engineering Electromagnetics", 2nd ed. Englewood Cliffs, New Jersey: Prentice-Hall, Inc, (1987).
- [16] A. M. Nicolson, G. F. Ross, "Measurement of the Intrinsic Properties of Materials by Time Domain Techniques", *Instrumentation and Measurement*, Vol. 19, (1970), pp.377-382., doi: 10.1109/TIM.1970.4313932.,
- [17] A.M. Noll, "Cepstrum Pitch Determination," *Journal of the Acoustical Society of America*, Vol. 41, No. 2 (1967) pp. 293-309.
- [18] J.G. Proakis, D. G. Manolakis, "Digital signal processing". Pearson Prentice Hall. ISBN 978-0-13-187374-2, (2007).
- [19] J.R. Goodwin, L.M. Hafner, P.M. Fredericks, " Raman Spectroscopy of the Heterogeneity of Microcolonies of a Pigmented Bacterium", *Journal of Raman Spectroscopy* 37 (2006) P.932-936.

# Feature Based Correspondence: A Comparative Study on Image Matching Algorithms

Usman Muhammad Babri<sup>1</sup>, Munim Tanvir<sup>2</sup> and Khurram Khurshid<sup>3</sup>

<sup>1,2,3</sup>Electrical Engineering Department, Institute of Space Technology  
1, Islamabad Highway, Islamabad, Pakistan

**Abstract**—Image matching and recognition are the crux of computer vision and have a major part to play in everyday lives. From industrial robots to surveillance cameras, from autonomous vehicles to medical imaging and from missile guidance to space exploration vehicles computer vision and hence image matching is embedded in our lives. This communication presents a comparative study on the prevalent matching algorithms, addressing their restrictions and providing a criterion to define the level of efficiency likely to be expected from an algorithm. The study includes the feature detection and matching techniques used by these prevalent algorithms to allow a deeper insight. The chief aim of the study is to deliver a source of comprehensive reference for the researchers involved in image matching, regardless of specific applications.

**Keywords**—computer vision; image matching; image recognition; algorithm comparison; feature detection

## I. INTRODUCTION

The recent decade has experienced drastic improvements in the field of computer vision, including scene or object recognition, stereo correspondence and motion tracking. Matching images and finding correspondence between them has been a key application to computer vision. This paper is a comparative study evaluating the performance of prevalent image matching algorithms. The performance is assessed according to the results obtained from various criteria such as speed, occlusion, sensitivity, etc. The paper also focuses on the level of accuracy achievable by an algorithm depending on the image type. This prediction ability comes in handy under the circumstances where resources are limited.

Achieving highly reliable results is the ultimate goal of any image matching method. However, none of them has gained universal acceptance. The type of image and the variations in the images to be matched are key elements in the selection of a matching method, along with the scale (two features of images have diverse scales), occlusion (two objects that are spatially separated in 3D plane might be interfering with each other when projected in 2D), orientation (two images are rotated with regard to each other), object to be matched (if the object is a planar, textured or edgy object), clutter (conditions of the image background) and illumination (fluctuations in illumination in two features).

The present image matching algorithms perform adequately under the above described conditions but even the most prevalent algorithms have not gained total invariance to these problems. However, through a methodical testing program, the

efficiency of the image matching algorithms will be tested on diverse sets of images with a noticeable difference in texture, clutter, orientation and other factors. These results will be eventually used to contrast between the distinctiveness of the found features. Therefore, our purpose here is to weigh the performance of these algorithms under various conditions using impartial standards.

## II. RELATED WORK

The roots of image matching and feature detection date back to 1981 to the work of Moravec using a corner detector for stereo matching. Initially, its applications were confined to stereo and short range motion tracking, but the revolutionary work of Schmid and Mohr (1997) illustrated how local invariant feature matching could be applied to image recognition problems for matching feature against a large database of images. They used Harris corner detector (1988) to select interest points which permitted features to be accorded under random orientation change. The Harris corner detector, however, was sensitive to changes in image scale. This was overcome by Lowe (1999) who achieved scale invariance by extending the local feature approach and then further extending it making the local features invariant to full affine transformations (Brown and Lowe, 2002).

### A. Scale Invariant Features Transform (SIFT)

Shortly after presenting an algorithm on feature detection from textured images, Lowe gave an improvised version of his work under the publication “Scale Invariant Feature Transform (SIFT) algorithm” (Lowe, 2004). This presents a technique to identify unique invariant features which may be used for comparison between different angles, orientations or viewpoints of an object or a scene. Lowe’s approach is broken down into four key components:

1) *Scale space extrema detection: to recognize probable interest points that are invariant to orientation and scale by means of a difference-of-Gaussian function.*

2) *Key point localization: a thorough model is used to define location and scale at every candidate location and then, centered upon measures of their stability, key points are selected.*

3) *Orientation assignment: according to local image gradient directions one or more orientations are allocated to every key point location. This provides invariance to changes in orientation, scale and location as all future calculations will be made relative to the assigned ones.*

4) *Key point descriptor*: at the area around each key point at the designated scale the local image gradients are measured and converted into a representation that tolerates substantial levels of variation in illumination and local shape distortion.

These stages are implemented in a descending order and only the key points that are robust enough jump to the subsequent stage. However, this proved to be an expensive process and therefore an upgraded version of SIFT's descriptor was presented "Principle Component Analysis SIFT (PCA-SIFT)" (Ke and Sukthankar, 2004). However, it proved to be less distinctive than SIFT. Another noteworthy approach is presented by Lindeberg under the name of "Scale Invariant Feature Transform" (Lindeberg, 2012). This technique for feature detection in the SIFT operator can be perceived as a variant of a scale-adaptive blob detection method where the detected blobs with related scale levels are found from scale-space extrema of the scale-normalized Laplacian.

### B. Speeded Up Robust Features (SURF)

Shortly after the PCA-SIFT another image matching algorithm was put forward that was to ensure speed in: detection, description and matching. This algorithm was named Speeded-Up Robust Features (SURF) detector (Bay et al., 2006). Contrary to other prevalent approaches of the time, SURF uses hessian matrix to considerably increase the matching speed. It depends on integral images to lessen the computation time and its descriptor defines a scatter of Haar-wavelet responses around the interest point. With the low dimensionality descriptor (64-dimensions) and 'Fast-Hessian' detector SURF is certain to perform faster.

### C. Features from Accelerated Segment Test (FAST)

Created by Trajkovic and Hedley (1998), FAST is the only feature-based algorithm used for this comparison. However, the implementation used for the comparison was published by Edward Rosten (Rosten and Drummond, 2006). The FAST detector is a wedge type detector i.e. a corner is detected using a circle surrounding a candidate pixel. It operates by considering a circle of 16 pixels and if there happen to be  $n$  adjoining pixels are above or below a threshold value,  $t$ , then the candidate pixel is chosen to be a corner. However, Rosten and Drummond extended this algorithm to use a machine learning based detector. Where all other detectors identify corners using an algorithm this technique trains a classifier on the model and then apply the classifier to an image. The classifier can be trained on how a corner should behave. This

makes the detector to perform significantly faster than other feature detectors.

### D. Other comparative studies

Various people have published studies to compare different image matching algorithms. Few of the most recent of these are Comparative Study of Image Matching Algorithms (Babbar et al., 2010), A comparison of SIFT, PCA-SIFT, and SURF (Juan and Gwon, 2009) and A Comparative Study of Three Image Matching Algorithms: Sift, Surf, and Fast (Guerrero, 2011). They have been effective in relating most aspects of the algorithms. However, their work does not present us either with the pictures they have used or with the system specifications or only textured images being used for the comparison.

The study aims at complementing the already available studies by allowing the quick recognition of the process that will perform efficiently under the given specifications. The paper also focuses on investigating the accuracy of the algorithms in detecting a single object in a scene.

## III. METHODOLOGY

The existing approaches towards the evaluation of the prevalent algorithms depend on their performance when applied on similar datasets. This has directed to different conclusions where one of these algorithms is presented as the finest, while in another publication the same algorithm performed contrarily. In our opinion, the major reason for this is that the image matching algorithms are best suited to a particular type of image and they would perform better when tried on these particular types of images. The suggested study will be based on the comparison of different image datasets. Herewith the hypothesis declaring that the performance of an algorithm depends on the type of image dataset.

Three image datasets will be taken under consideration during the study to compare the image matching algorithms. The dataset to be used is as follows:

- 1) *An image with repetitive patterns*
- 2) *An image with a cluttered background*
- 3) *An image of a planar object*

These images are pictures taken from an Infinix hot note's rear mobile camera with a resolution of 8MP. Altogether the images correspond to indoor and outdoor conditions. The original images used were 3840x2160. However, these images were too large for the algorithms and hence were resized by a scale factor of 1/8 for efficient processing.

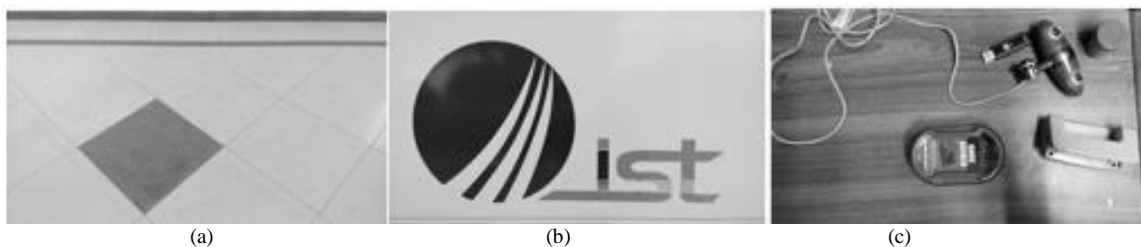


Fig. 1. Floor tiles at the Institute of Space technology (b) A logo of Institute of Space Technology (c) A shoe shiner in a cluttered scene



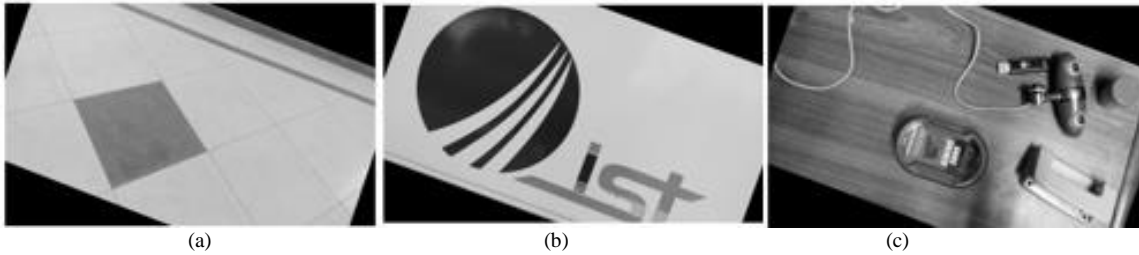


Fig. 2. (a) Floor tiles at the Institute of Space technology rotated at an angle of  $30^{\circ}$  (b) A logo of Institute of Space Technology rotated at an angle of  $30^{\circ}$  (c) A shoe shiner in a cluttered scene rotated at an angle of  $30^{\circ}$

Three of the most prevalent matching algorithms that are to be used for the study are SIFT, SURF and FAST.

SIFT is perhaps the most prevalent image matching algorithm. It can equate images under various scales, orientations and illuminations. However, it was found to be substantially slow. Various implementations for the code are available on the web. For this study the implementation being used was initially created by D. Alvaro and J.J. Guerrero, Universidad de Zaragoza and then modified by D. Lowe.

SURF, published after SIFT, was proposed to overcome the shortcomings of SIFT which include the computational cost and the execution time. The SURF implementation used in this study is a direct translation of the Open Surf C# code of Chris Evans.

FAST, first developed by Trajkovic and Hedley in 1998, is the lone feature based algorithm being used for the study. The implementation used was published by Edward Rosten (Rosten and Drummond, 2009). However, a matching module was not available for FAST as a MATLAB implementation. Though a bit unfair but a proportionality analysis could be performed to help keep it as fair as possible. All these implementation are tested using MATLAB R2015a on an Acer Aspire V5 with a 1.9GHz processor and 4.00GB RAM.

#### IV. EXPERIMENTATION

Each algorithm will be used to test the images shown above and these, consequently, will be tested in diverse categories with the final and chief goal of the evaluation of the intrinsic algorithm features that differentiate it from others. Parameters such as speed, number of feature and number of matches will be assessed. The images in this experiment will be taken in pairs with a rotation less than  $30^{\circ}$  from the image to its corresponding pair. A rotation of 30 degrees is chosen as it is the typical maximum value for most of the algorithms to perform a reliable research.

Once the match pairs are selected, the algorithms are applied to each image pair to quantify the number of features detected by each of the algorithm. However, the number of features detected is not a good measure of performance by itself because detecting 10 important features is better than detecting 10 features that have a lesser chance of finding a match pair. Therefore, a visual scrutiny was performed to manually match 10 features per image pair. Then these features were checked for a match by applying the algorithms to the images. The obtained results were then tested for the number of errors sustained. Type I and Type II error classification was

used for the process where Type I errors are the ones in which real matches are not sensed by the algorithms and Type II are the ones in which a feature is mismatched. The efficiency of the algorithm is calculated using the relation:

$$\%Efficiency = \frac{2 \times \text{No.of matches}}{\text{Total No.of features detected}}$$

The efficiency of each algorithm are presented in table II, III and IV. The results show that the number of matches are significantly lesser than the number of features detected. Therefore, it establishes that the number of matches are by themselves not very good. What is needed is to compare that how much of features that are detected by the algorithm are actually useful in the process of matching them to their correspondents.

#### V. RESULTS

The experimental follow up produced some likely results especially with the feature detection components of the algorithms. SURF detects better quality features than SIFT, in general, as SIFT fails to match a large number of features that it detects.

It is vital to remember that the analysis does not compare the types of images most suited to the algorithms. This can be observed in the following results as SIFT and SURF both being textured based algorithm present similar results on same images.

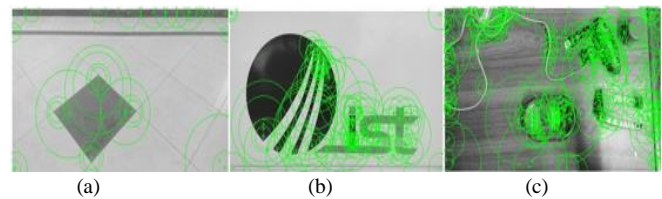


Fig. 3. SIFT feature detection on (a) Floor tiles at the Institute of Space technology (b) A logo of Institute of Space Technology (c) A shoe shiner in a cluttered scene

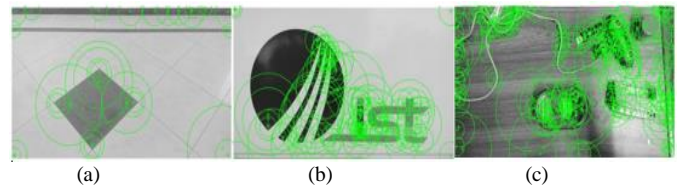


Fig. 4. SURF feature detection on (a) Floor tiles at the Institute of Space technology (b) A logo of Institute of Space Technology (c) A shoe shiner in a cluttered scene

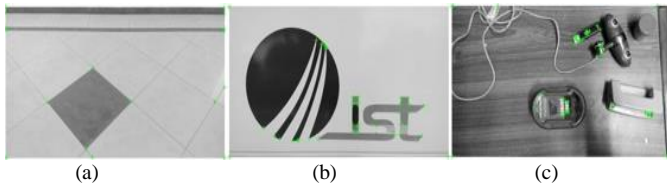


Fig. 5. FAST feature detection on (a) Floor tiles at the Institute of Space technology (b) A logo of Institute of Space Technology (c) A shoe shiner in a cluttered scene

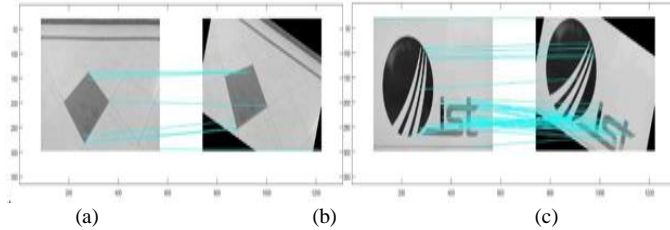


Fig. 6. SIFT matching on (a) Floor tiles at the Institute of Space technology (b) A logo of Institute of Space Technology

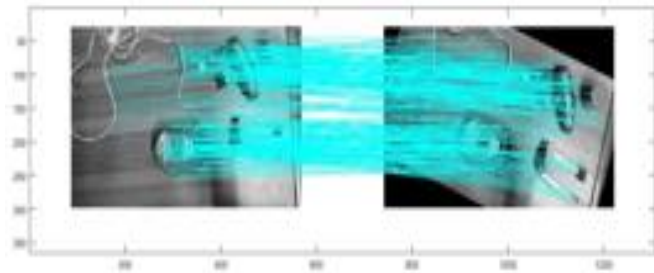


Fig. 7. SIFT matching on a shoe shiner in a cluttered scene



Fig. 8. SURF matching on (a) Floor tiles at the Institute of Space technology (b) A logo of Institute of Space Technology (c) A shoe shiner in a cluttered scene

TABLE I. FEATURES DETECTION USING SIFT, SURF AND FAST

Features Detected				
Pair	Image	SIFT	SURF	FAST
Floor	A	209	117	17
	B	129	93	10
IST logo	A	333	171	97
	B	162	83	112
Cluttered scene	A	675	552	47
	B	642	481	32

TABLE II. FEATURES MATCHING AND MATCHING EFFICIENCY USING SIFT, SURF AND FAST

	SIFT		
	Features	Matches	%Efficiency
Floor	209	19	11.24
	129		
IST logo	333	50	51.28
	162		
Cluttered scene	675	226	34.32
	642		
	SURF		
	Features	Matches	%Efficiency
Floor	117	29	27.62
	93		
IST logo	171	42	33.07
	83		
Cluttered scene	552	130	25.17
	481		
	FAST		
	Features	Matches	%Efficiency
Floor	17	9	58.06
	14		
IST logo	97	26	24.88
	112		
Cluttered scene	47	22	55.7
	32		

## VI. CONCLUSIONS

Given that SIFT is an established robust feature detector, it is not surprising that a large amount of features are detected in the images. SIFT detects a large amount of features in textured images as evident from the results. However, a lot of features it detected were in places with insufficient information for matching it later. But still the high number of features detected compensate for the features in areas with less information.

Following SIFT, SURF also detects a high number of features. But the difference between the features detected by each of the algorithm is significant. SURF also being a texture-based algorithm works best on textured images but doesn't give a lot of features with planar ones. Although SURF finds less features, it finds more features in areas with more information as compared to SIFT. This gives a positive sign about the algorithm as the features detected by the algorithm are more likely to be matched.

FAST, in contrast to its competitors, detects limited features. Being a feature based algorithm this could be expected of the algorithm. But such algorithms limit themselves to detect features with high contrast to the feature surroundings. Therefore, the features detected were among the most robust ones. FAST works best with planar images and the quickest too. This proves our hypothesis that the algorithm used depends on the type of image being used.



## VII. FUTURE CONTRIVES

Future prospects of this research include testing the accuracy of image matching algorithms in detecting a single object in a scene and an in-depth analysis of the algorithm structure to perform an automated assessment of the algorithm's performance in various scenarios.

### REFERENCES

- [1] Babbar, G., Punam Bajaj, Anu Chawla, and Monika Gogna. A comparative study of image matching algorithms. *International Journal of Information, Technology and Knowledge Management*. July-December. 2(2): 337-339. 2010.
- [2] Bay, H., A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision ECCV 2006*, Vol. 3951. *Lecture Notes in Computer Science*. p. 404-417. 2006
- [3] Dorkó, G., and C. Schmid. Selection of scale-Invariant parts for object class recognition. *Proceedings on Ninth IEEE International Conference on Computer Vision 2003*, 13-16 Oct. 2003, Nice, France. 2003.
- [4] Harris, C. A combined corner and edge detector. *Fourth Alvey Vision Conference*, Manchester, UK. p. 147-151. 1998.
- [5] Juan, L., and O. Gwun. A comparison of SIFT, PCA-SIFT, and SURF. *International Journal of Image Processing (IJIP)* 3(4): 143-152. 2009.
- [6] Ke, Y., and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)* – 2: 506-513. 2004.
- [7] Lowe, D.G. Object recognition from local scale-invariant features. *The proceeding of the seventh IEEE International Conference on Computer Vision 1999*. 2: 1150-1157. 1999.
- [8] Lowe, D.G. Distintive image feature from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2): 91-110. 2004.
- [9] Mikolajczyk, K, and. C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision* 60(1): 63-86. 2004.
- [10] Trajkovic, M. and M. Hedley. FAST corner detector. *Image and Vision Computing* 16: 75-87. 1998.
- [11] Guerrero, Maridalia. A Comparative Study of Three Image Matching Algorithms: Sift, Surf, and Fast. *All Graduate Theses and Dissertations*. Paper 1040. 2011.
- [12] Rosten, E., Drummond, T. 2005. Fusing points and lines for high performance tracking. In *10th IEEE International Conference on Computer Vision*. Volume 2., Beijing, China, Springer. 1508-1515
- [13] Rosten, E., and T. Drummond. Machine learning for high-speed corner detection, *IEEE Trans. Pattern Analysis and Machine Intelligence* 2006. p. 430-443. 2006.
- [14] Lindeberg, T. Scale invariant feature transform. *Scholarpedia* 7(5), 10491. 2012
- [15] B. Pinto and P. R. Anurenjan, "Video stabilization using Speeded Up Robust Features," *Communications and Signal Processing (ICCSPP)*, 2011 International Conference on, Calicut, pp. 527-531.
- [16] Elmar Mair, Gregory D. Hager, Darius Burschka, Michael Suppa and Gerhard Hirzinger. Adaptive and Generic Corner Detection Based on the Accelerated Segment Test. *Computer Vision ECCV 2010*, Vol. 6312. *Lecture Notes in Computer Science*. p. 183-196. 2010.

# New Mathematical Modeling of Three-Level Supply Chain with Multiple Transportation Vehicles and Different Manufacturers

Amir Sadeghi \*

PhD Student, Department of  
Industrial Management,  
College of Management and  
Accounting,  
South Tehran Branch, Islamic Azad  
University,  
Tehran, Iran

Amir Farmahini Farahani

PhD Student, Department of  
Industrial Engineering,  
South Tehran Branch, Islamic Azad  
University,  
Tehran,  
Iran

Hossein Beiki

Faculty Member, Department of  
Industrial Engineering,  
Naragh Branch, Islamic Azad  
University,  
Naragh,  
Iran

**Abstract**—Nowadays, no industry can move in global markets individually and independently of competitors because they are part of a supply chain and the success of each member of the chain influences the others. In this paper, the issue of three-level supply chain with several products, one manufacturer, one distributor, and several customers is reviewed. In the first part of the chain, one type of vehicle and in the second part of the chain, two types of vehicles are used. The proposed model of this paper is a mixed integer planning integrated model. The model is considered to minimize costs, including transportation, inventory, and shortage penalty cost. This paper is about the proposed approach for development of quantitative models in the field of three-level supply chain and theoretical research and presents a case study of sending produced rolls by Mobarakeh Steel Structure Company to “Sazeh Gostar Saipa (S.G.S)” and then to suppliers. The solution proposed in this paper is imperialist competitive algorithm which is solved in 20 different sizes, and the results in small size are compared with GAMS.

**Keywords**—Transportation; Mathematical Model; Logistic Costs; Imperialist Competitive Algorithm

## I. INTRODUCTION

In a supply chain, transportation is an activity that influences all elements from the beginning of production until the delivery of products to customers; and in all activities, an efficient transportation significantly influences the reduction and augmentation of costs. Given the importance of a supply chain, this study attempts to propose a mathematical model for a three-level supply chain with the aim of minimizing chain costs so that products (rolls produced by Mobarakeh Steel Company) are transported from Mobarakeh Steel Company to Sazeh Gostar Saipa Company using one type of vehicle and then from Sazeh Gostar Saipa to suppliers using two vehicles. The proposed model for this research is a mixed zero and one planning integrated model. The outputs of the model include the number of rolls transported from manufacturer to Sazeh Gostar Company and then from Sazeh Gostar to each supplier, roll shortage for suppliers, inventory level of rolls in stock at the end of each period, and departure time of vehicles in the period. This model is solved in 20 different sizes. To solve the

problem in small and large sizes, GAMS software and imperialist competitive algorithm are used, respectively; and imperialist competitive answers are found to be superior. Each problem has been implemented five times, and the best solution (least cost) is obtained and then the standard deviation has been examined five times.

The overall outline of this paper is as follows: initially a brief review of literature is provided, then the mathematical model is presented in Section 2, and after evaluation solutions in Section 3. Section 4 shows numerical examples and results, conclusion is presented in Section 5. Boumool and Vinod [1] conducted the first attempts on planning transportation and inventory. They proposed a theoretical model that could determine the number of orders and transportation options at the same time. The purpose of this model was to minimize the total cost of transportation, ordering, and inventory maintenance. The model considers a fixed cost to carry each item. Yukuyuma [2] has proposed an optimization integrated model with a possible demand in which inventory control and transportation are taken into consideration simultaneously. Products of a distribution center are transported to several customers. He has proposed a method to locate re-order point, determine the amount of re-order, and the number of transported products so that the total cost of the inventory and transportation will be minimized. Kim [3] has reviewed the schedule of sending in a system consisting of a central stock and  $n$  retailers. In the studied transportation system, only one type of product is taken into consideration. The purpose of this paper is to minimize the inventory cost of retailers so that they do not face any shortage.

Asgari and Aghdasi [4] have proposed an integrated mathematical model to plan inventory and transportation decisions. This multi-objective model describes minimizing maintenance costs and buyer orders, minimizing transportation cost of manufacturer, minimizing possibility of delay in the delivery of parts to achieve timely delivery and maximum utilization transportation fleet capacity. The model is intended for several products of chain; and if the demand for products is certain, several types of vehicles are also used to send products.

Young and Beyung Park [5] have examined scheduling issue of vehicles taking the service and delivery time into account to minimize the total travel time and product delivery delay to retailers. This model is solved with hybrid genetic algorithm and its objective function is mixed integer programming. Karabook [6] examined the allocation of tasks to manufacturers with the aim of determining optimal production plan so that work completion time is minimal. Different manufacturers may need different time to process each job. To solve this problem, an adopted genetic algorithm, which is a new mechanism called gene mutation was applied.

Chue, Cheng, and Lin [7] have studied different ways to minimize work completion time on a two-level supply chain. They postulated that there are two manufacturers in parallel form on the first level and there is a factory on the second level. Things come in certain groups and researchers assumed that a fixed setup time is required for each group. Zegardi [8], considered the production and scheduling of transportation in a two-level supply chain in this study. Its objective function minimizes the total delay and total deviation from work load contributed to manufacturers from shares allocated to them. Its mathematical model is mixed integer programming. MSGA is applied to solve this problem and it is also compared with genetic algorithm and the results show the superiority of MSGA to GA. In this study, it is assumed that manufacturers are located in many geographical areas. Numerical examples of this study consist of 6 works, 3 manufacturers, and 2 vehicles.

David Naso's [9] paper is a combination of timely production and transportation, indicating that timely delivery to customers is one of the most important aspects of supply chain management. This article assumes that outsource is sometimes necessary for production. In this research, both earliness and delay are prohibited and genetic algorithm was used to solve it. It has a two-level supply chain, its objective function aims to minimize costs including outsource cost, waiting time punishment cost, and transportation cost. Zandiyeh *et al.*, [10] hypothesized that an unlimited number of sources are available to deliver completed works without the need for a long time to move them. Thus, in this article, the production and scheduling of air transportation has been considered at the same time. The whole problem consists of two sub-problems which are air transportation and scheduling of a vehicle taken into consideration simultaneously. Objective function minimizes the entire chain costs and the time needed to do things such as earliness costs, departure delay costs, and move costs. Ghasemi *et al.*, [11] designed a mathematical model for inventory management by programming with multiple objectives, developing an objective multi-function mathematical model is the certain finding of this study whose input is supplier annual demand, production capacity of manufacturer, type of piece packing, etc. and its output includes, production volumes (the amount of orders), the amount of storage required for each partner at central stocks, scheduling of sent batch from partners, etc. where reducing transport costs, optimizing the times of sending shipments from suppliers, and reducing the number of received shipments in the parent company is taken into consideration; and ultimately regarding statistical population of the study which is Saipa Automotive Group, developed mathematical model is tested and its output is

compared with actual values of Saipa Company. Ann Zhou *et al.*, [12] stated that inventory, routing, and scheduling are three main factors for the operation of supply chain. They reviewed inventory integrated model, routing, and scheduling in a two-level supply chain, VNS algorithm is used to solve NP problem of their study; then the results of this algorithm is compared with those of other algorithms and the superiority of this algorithm is approved. In this model, there is only one supplier and each route starts from one supplier and ends up to this very same supplier.

Vehicle capacity is already identified. Objective function of this paper aims to minimize inventory and transportation costs. Gorouprasad Pundur [13], in this article, scheduling of production and distribution in two-level supply chain including a number of manufacturers, one stock and customer is taken into consideration. Each manufacturer produces various products with fixed production rate; customer demand is constant for each product at all times; and exact and heuristic methods are applied and compared to solve the model. The aim of objective function is to reduce costs, including inventory cost, distribution costs, and set-up cost. The model is of mixed integer non-linear programming model type. Behatnagar, Mehta, and Theo [14] tried to balance decision problems of programming and short-term scheduling in supply chain using dual models. Two transporting methods are presented in this article one of which is maritime transportation with longer delivery time and less delivery cost, and another one is air transportation with more delivery cost, less delivery time, and more accurate prediction of demand. Decisions of programming determine the amount of order and inventory in maritime transportation, while decisions of scheduling concern scheduling and amount of air transportation. Numerical results of the study show that this model leads to sustainable cost recovery in a wide range of operations of both methods.

Alyamam and Dedin [15] modeled integrated supply chain as a network project, this model balances stationary and mobile inventory maintenance costs, operation cost in less time, transportation cost and delay penalty cost in the delivery of customer orders.

Hisham al-Din, Serker, and Aisam [16] investigated a two-step retrieved model to produce an inventory system with the possibility of damage in transportation. The aim of this model is to minimize related costs. It has a heuristic solution that deviates slightly from accurate methods. The results indicate that optimizing scheduling depends heavily on the relation between parameters of lost sales cost and timely sales cost.

Osman and Demireli [17] studied accumulated economic problems and delivery scheduling for multi-stage supply chain, and investigated maximum inventory strategy, and identifying throughput and time cycle for maximum inventory simultaneously so that all costs are minimized. These two methods are combined by non-linear models with a hybrid algorithm. Based on numerical results of the presented model, given algorithm has obtained an optimal solution for large-scale problems in a short time, thereby reducing costs by 16.3%.

In Kim *et al.*, paper [18], two models are proposed to minimize total costs of transportation and inventory

maintenance in a two-level supply chain in which dynamic demand and product transportation take place from a distribution center to customers by one type of vehicle. Transportation costs include fixed transportation cost and fixed handling costs of materials related to customer. Their proposed model is intended only for one type of product, but the necessity of expanding this model in following areas was realized due to the variety of products in reality:

- In proposed model, transportation cost consists of the times of transportation.
- In proposed model, distributor is evaluated as well as manufacturers and customers.
- Several products (rolls with different dimensions) are used instead of one type of product.
- To send products from manufacturer to distributor, one kind of vehicle is used; and to send products from distributor to customers, two types of vehicles are used.

In an article by Zegardy *et al*, [19] production and transportation scheduling have been studied. Several vehicles with different transportation capacity and speed are used. This article assumes that all suppliers are in one location. It is extended in following ways:

- One supplier is used instead of several suppliers.
- Instead of minimizing the time to do things, costs are minimized.

## II. MATHEMATICAL MODEL

In this section, scheduling model of transportation and inventory is proposed to minimize total cost of transportation, inventory and shortage penalty. First, applied assumptions and symbols are expressed:

### A. Assumptions

- In the first part of chain, one type of vehicle is used to move goods, and in the second part of chain two types of vehicles with recognized transportation capacity are used.
- This model is intended for a product with different dimensions.
- Shortage is allowed and is in the form of lost sales.
- At the beginning of scheduling, vehicles are available and rolls will be transported immediately.
- Unloading and storage time is clear for vehicle.
- Demands of a supplier are met in each travel.
- Demand rate of a supplier is supposed to be clear for the manufacturer.
- Transportation cost only includes the cost of the times of transportation and it is different for each manufacturer according to distance.
- A 20-ton truck trailer is the vehicle from Mobarakeh

Steel Company to Sazeh Gostar and a 20-ton or a 10-ton truck trailer can be the vehicle from Sazeh Gostar to suppliers.

- Only one type of 20-ton truck trailer is used.
- Production programme has been almost constant for any desired periods (weekly, monthly, and quarterly) through leveling.
- Storage at supplier is impossible, and demand for period  $t$  cannot be sent during earlier periods.
- A vehicle can move one time in period  $t$  from factory to stock or from stock to supplier.

### B. Indices

I. Sum of all products (rolls with different dimensions)

i. Product index

J. Set of all suppliers (customers)

j. supplier index ( $j \in J$ )

K. Set of all vehicles from manufacturer to stock

k. vehicle index

$K'$ . Set of all vehicles from stock to supplier

$k'$ . Vehicle index ( $k' \in K'$ )

T. Set of all periods

t. Period index ( $t \in T$ )

### C. Parameters

$a_k$  = fixed transportation cost between manufacturer and stock by vehicle  $k$

$b_{jk'}$  = fixed transportation cost between stock and supplier  $j$  by vehicle  $k'$

$h_{it}$  = maintenance costs per unit of product (roll  $i$ ) by stock in period  $t$

$r_i$  = shortage penalty per unit of product (Roll  $i$ ) in stock

$D_k$  = capacity of vehicle  $k$  according to weight to travel from manufacturer to stock

$D'_{k'}$  = capacity of vehicle  $k'$  according to weight to travel from stock to supplier

$g$  = stock capacity

$w_i$  = weight of each packet of product  $i$

$d_{ijt}$  = product demand  $I$  for supplier  $j$  in period  $t$

$ct_k$  = travel time from manufacturer to stock by vehicle  $k$

$ct'_{k'j}$  = travel time from stock to supplier  $j$  by vehicle  $k'$

$st_t$  = beginning of period  $t$

$Fin_t$  = end of period  $t$

D. Variables

$X_{ikt}$  = number of product (roll i) transported from manufacturer to stock by vehicle k in period t

$Q_{ijk't}$  = number of product (roll i) transported from stock to the supplier j by vehicle k' in period t

$S_{ijt}$  = number of product shortage (roll i) for a supplier j in period t

$In_{it}$  = number of product inventory i in stock at the end of period t

$Z_{kt}$  = Is taken 1 if vehicle K is sent to stock in period t, otherwise is taken zero.

$Z'_{jk't}$  = Is taken 1 if vehicle k' works for supplier j in period t, otherwise is taken zero.

$c_{kt}$  = departure time of vehicle K from manufacturer to stock in period t

$c'_{jk't}$  = departure time of vehicle k' from stock to supplier j in period t

E. Formulating the problem mathematically

Now, according to mentioned assumptions, symbols, and variables, mathematical model of problem is as follows:

Min Z

$$\sum_t \sum_k a_k (z_{kt}) + \sum_t \sum_j \sum_{k'} b_{jk'} (z'_{jk't}) + \sum_t \sum_i h_{it} (In_{it}) + \sum_i \sum_t \sum_j r_i s_{ijt} \quad (1)$$

The aim of proposed mathematical model is to minimize total costs of transporting products from manufacturer to stock and from stock to customers, cost of inventory maintenance at the end of each period, and cost of product shortage penalty.

The first and second parts of objective function show transportation cost which is the product of fixed transportation cost in a zero and one variable indicating whether or not the given vehicle is used in each period.

The third part of the objective function represents total costs of inventory maintenance of products available in stock in different periods; the final part also reflects shortage penalty of products facing shortage in different periods.

s.t.

$$z_{kt} \geq \frac{\sum_i w_i x_{ikt}}{D_k} \quad \forall k, t \quad (2)$$

Limitation (2) shows the number of vehicle k used in each period to transport products from Mobarakeh Steel to Sazeh Gostar stock; in addition, it shows the low limit of the capacity of each vehicle.

$$z'_{jk't} \geq \frac{\sum_i w_i Q_{ijk't}}{D'_{k'}} \quad \forall k', t, j \quad (3)$$

Limitation (3) indicates the number of vehicles k' used in each period to transport products from Sazeh Gostar stock to suppliers; in addition, it indicates the low limit of the capacity of each vehicle.

$$\sum_k x_{ikt} + In_{i,(t-1)} - \sum_{k'} \sum_j Q_{ijk't} = In_{i,t} \quad \forall t, i \quad (4)$$

Limitation (4) represents stock inventory variable. Product inventory i in stock at the end of period equals the number of products sent to Sazeh Gostar stock in different periods by available vehicles, as well as product inventory i at the end of previous period less the number of product which is sent to supplier by all available vehicles.

$$d_{ijt} = \sum_{k'} Q_{ijk't} + s_{ijt} \quad \forall i, j, t \quad (5)$$

Limitation (5) shows the demand of each supplier is met; the demand of each supplier in each period equals the number of products sent to each supplier in different periods by available vehicles, as well as shortage of that product for each supplier.

$$\sum_i w_i In_{it} \leq g \quad \forall t \quad (6)$$

Limitation (6) represents stock capacity limitation. The volume of products stored in stock shall not exceed the useful capacity of stock. Left side of limitation shows the product of volume of products and their amount of inventory and right side manifests stock capacity.

$$\sum_i x_{ikt} \leq MZ_{kt} \quad \forall k, t \quad (7)$$

Limitation (7) shows the relationship between two variables xikt and zkt .

$$\sum_i Q_{ijk't} \leq MZ'_{jk't} \quad \forall k', t, j \quad (8)$$

Limitation (8) shows the relationship between two variables  $Q_{ijk't}$  and  $Z'_{jk't}$  .

$$c_{kt} \geq c_{k,t-1} + 2ct_k + T + 5 - m(1 - Z_{k,t-1}) \quad \forall k, t > 1 \quad (9)$$

Limitation (9) represents departure time of vehicle k to stock. This limitation shows that departure time of vehicle k to stock in each period is equal to departure time of each vehicle in previous period, as well as the sweep time of vehicle plus the time it takes to unload the vehicle in stock as well as rest periods for drivers.

$$c'_{jk't} \geq c'_{jk'(t-1)} + 2ct'_{jk'} + T' + 5 - m(1 - Z'_{jk'(t-1)}) \quad \forall j, k', t > 1 \quad (10)$$

Limitation (10) represents departure time of vehicle k' to suppliers. This limitation shows that departure time of each vehicle to stock in each period is equal to departure time of each vehicle in previous period, as well as the sweep time of vehicle plus the time it takes to unload the vehicle in stock as well as rest periods for drivers.

$$st_t z_{kt} \leq c_{kt} \leq Fin_t z_{kt} \quad \forall k, t \quad (11)$$

$$st_t z'_{jk't} \leq c'_{jk't} \leq Fin_t z'_{jk't} \quad \forall j, k', t \quad (12)$$

Limitations (11) and (12) show that departure time of each vehicle to the destination in the period must be between the beginning and the end of the period.

$$\sum_j Z'_{jk't} \leq 1 \quad \forall t, k' \quad (13)$$

Limitation (13) indicates that each vehicle is sent to one customer and the demand of one supplier can be met in each travel.

$$c_{kt}, c'_{jk't} \geq 0 \quad (14)$$

$$x_{ikt}, Q_{ijk't}, In_{it}, S_{ijt} = \text{Integer} \quad (15)$$

$$Z_{kt}, z'_{jk't} \in (0, 1) \quad (16)$$

Equation (14) indicates that departure time variables of vehicle from manufacturer to stock and departure time of vehicle from stock to suppliers are positive temporarily. Equation (15) also shows that variables of the number of products transported from manufacturer to stock and the number of transported products from stock to suppliers as well as the number of product shortage and inventory level are all integer variables. And relation (16) also indicates that the variable of use or non-use of any vehicle in each period is as zero and one.

### III. SOLUTION

Like other evolutionary algorithms, this algorithm begins with some random initial population, and each of them is called a country. Some of the best elements of population (equivalent elite in GA) elected as imperialist. The remaining population is considered the colony. Depending on their power, imperialists attract the colonies with a special process. Total power of the empire depends on both forming parts that are imperialist country (the core) and its colonies. Mathematically, this dependency is modeled by defining imperial power as total power of imperialist country, plus a percentage of the average of its colonial power. Once early empires are formed, imperial competition will start among them. Each empire that fails to compete in colonial competition and increase its power (or at least fails to prevent its influence reduction) will be removed from colonial competition. Thus, the survival of an empire depends on its ability to attract and dominate the colonies of rival empires. As a result, during imperialist competitions, the power of larger empires gradually increases, and weaker empires will be removed. Empires are forced to develop their colonies to increase their own power. Over time, the colonies will approach empires in power and you see some kind of a convergence of power. At the final stage of imperialist competition, there is only one united empire in the world with the colonies whose position is very close to the imperialist country.

#### A. Proposed pseudo-code Algorithm

- 1) Select some random points on the function and form up early empires.
- 2) Move the colonies to the imperialist countries (matching policies)
- 3) If there is a colony of an empire whose cost is less than the imperialist country, change the colonial and imperialist position.

4) Count total cost of an empire (regarding the imperialists and their colonies costs).

5) Select a colony from the weakest empire and give it to the empire who most likely dominates it.

6) Remove weak empires.

7) If only one empire remains, stop and go back to 2.

#### B. Parameter adjustment

As can be seen, the performance of an algorithm depends heavily on the intensity of its parameters. In this study, response surface methodology (RSM) is used to set parameters. Different levels of parameters can be coded as follows:

$$X_i = \frac{r_i - \left(\frac{h+l}{2}\right)}{\left(\frac{h-l}{2}\right)} \quad (17)$$

Where  $h$  and  $l$  denote upper and lower surfaces of parameters, respectively; and  $X_i$  and  $r_i$  denote coded and actual values of parameters, respectively. At this point, the algorithm will be administered for different combinations of parameters and their different levels. Another input of RSM methodology is an index that can compare the algorithm for different combinations of parameters. In this study, the same value of objective function is used as a response. Adjusted values of parameters are shown in Table 1.

TABLE I. ADJUSTED PARAMETERS IN THE PROPOSED ALGORITHM

Factors	Optimal coded value	Optimal real value
$n\text{-Pop}$	0.85	193
$N\text{-imp}$	-0.2	5
$P_A$	0.18	0.54
$P_R$	-0.8	0.12
$E$	0.9	0.195
$B$	-0.2	1.8

#### C. Numerical example

To test the algorithm, 20 example problems are produced by Sazeh Gostar Company regarding given numerical range. The data are explained in Table 2.

#### D. Solving the model

GAMS 22.1 software is used to code and solve the model, computer specifications are 2.2 GHZ, RAM 4, which have been reported in Table 3. The first column of the table indicates the number of problems which are 20. The second column shows the number of suppliers that is considered from 10 to 100. The third column indicates the number of products which is considered from 5 up to 70.

The fourth column shows obtained objective function values using GAMS software. Only 5 first samples had feasible answers, and for larger-scale problems, the imperialist competitive algorithm is used. Each of these samples was performed on a total of 5 times with the fifth column represents the best answer which had the least cost. The sixth column shows the standard deviation of 5 times of administration of each problem. The seventh column manifests computational problem solving time. The last column shows GAP which has been increased to 1.1.



TABLE II. PROBLEM DATA

Number of products	Number of customers	Sample issue
1	10	5
2	10	10
3	10	20
4	20	10
5	20	20
6	20	25
7	30	10
8	30	20
9	30	25
10	40	20
11	40	25
12	40	30
13	70	30
14	70	35
15	70	40
16	100	30
17	100	40
18	100	50
19	100	60
20	100	70

TABLE III. RESULTS OF THE MODEL

Pro. No.	Sup. No	Sam. issue	OFV of GAMS	OFV of ICA		Cpu Time (sec)	Gap
				Mean	Std		
1	10	5	3693276	3693276	50	15	0
2	10	10	8792932	8836897	63	25	0.5
3	10	20	95951810	96815376	98	35	0.9
4	20	10	138442170	140241918	79	49	1.3
5	20	20	1544783680	1561776300	101	78	1.1
6	20	25	-	3901481030	97	90	-
7	30	10	-	10721125650	110	110	-
8	30	20	-	18455602650	124	132	-
9	30	25	-	54754476050	132	156	-
10	40	20	-	166157490090	109	230	-
11	40	25	-	321982554408	107	276	-
12	40	30	-	643457896800	147	290	-
13	70	30	-	$7.6 \times 10^{15}$	152	331	-
14	70	35	-	$8.2 \times 10^{18}$	132	359	-
15	70	40	-	$5.3 \times 10^{22}$	126	380	-
16	100	30	-	$4.9 \times 10^{25}$	128	690	-
17	100	40	-	$3.3 \times 10^{27}$	113	701	-
18	100	50	-	$2.8 \times 10^{30}$	143	725	-
19	100	60	-	$7.4 \times 10^{35}$	179	758	-
20	100	70	-	$6.9 \times 10^{41}$	168	790	-

E. Graphs

Convergence trend of ICA is shown in the following figures. Vertical column of these graphs represents the best cost and horizontal column shows the number of iteration. It is observed in all graphs that the more developed is the graph, the better are the results.

Figure 1 is for problem number 7 with 30 suppliers and 10 products; and figure 2 is for problem number 20 with 100 suppliers and 70 products.

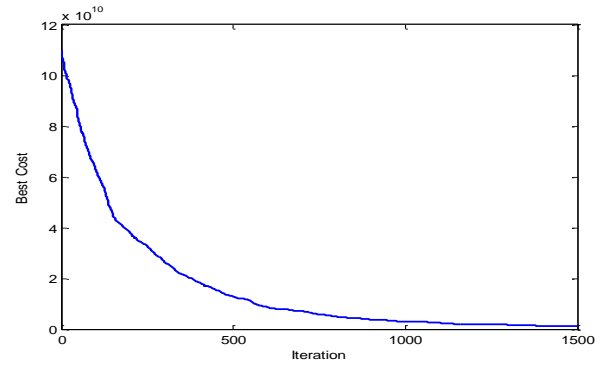


Fig. 1. The relationship between time and cost for problem number 7

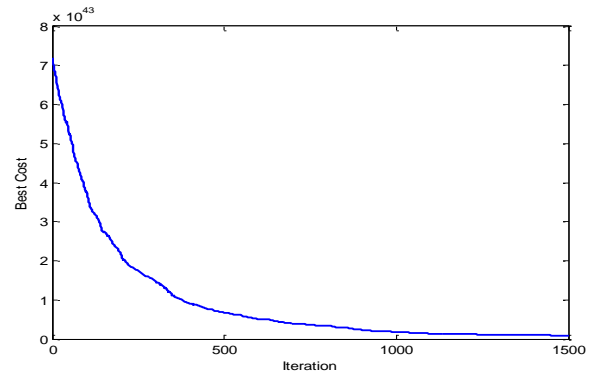


Fig. 2. The relationship between time and cost for problem number 20

F. Sensitivity analysis

In this section, sensitivity analysis has been used to verify the correctness of modeling and investigate to which parameter changes, objective function is more sensitive. Sensitivity analysis is conducted on three parameters that are the number of suppliers, the number of products, and demand; and the results are shown in following figures in which horizontal axis is for studied variable and vertical axis is for objective function. Figure 3 shows the changes in objective function vs. changes in the number of suppliers, the number of suppliers has been changed from 10 to 100, and objective function value is calculated for each of them.

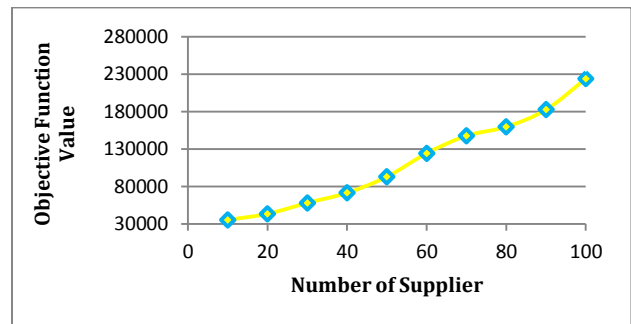


Fig. 3. Changes in objective function value vs changes in the number of suppliers

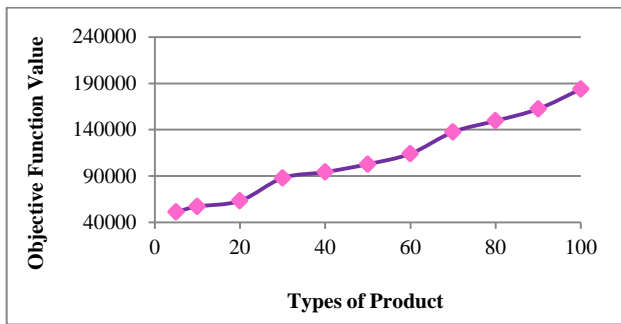


Fig. 4. Changes in objective function value vs. changes in the number of products

Figure 4 shows changes in the objective function value vs. changes in the number of products. The number of products has been increased from 5 to 100 and objective function value is also shown in figure.

Figure 5 shows changes in objective function value vs. changes in demand. Objective function value has increased following an increase in demand. The effect of this parameter on objective function was more significant than any other parameter; therefore, objective function is more sensitive to demand.

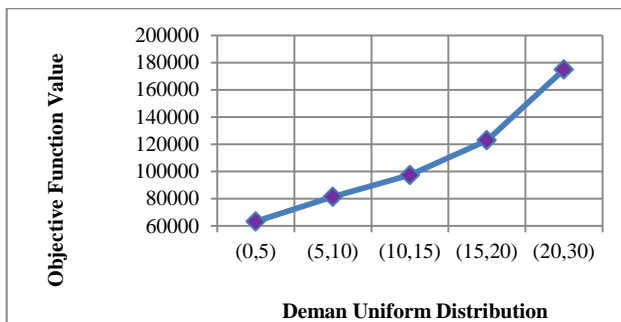


Fig. 5. Changes in objective function value vs. changes in demand

#### IV. CONCLUSION

In this paper, first mathematical model due to the restrictions of Sazeh Gostar Saipa (S.G.S) Company has been studied. Then a solution is proposed for this model which follows the imperialist competitive algorithm. It is then compared on a small scale with Gams and its superiority is verified. The distribution of input data, parameter set, result tables, and graphs is demonstrated and 20 numerical samples are presented to further the understanding of the problem. Since GAPs are less than 6%, and responses are obtained in a reasonable time, it can be concluded that imperialist competitive algorithm has been true. Then to verify the correctness of model, sensitivity analysis has been conducted on three parameters, and finally the results show that the model has demonstrated a great sensitivity to demand. Future research could be present developing mathematical model with greater levels of supply chain and with the more different number of vehicles and producers or distributors and comprehensive

development model suggested the results with the different algorithms and analyze, and also a multi-objective considering several important objectives simultaneously to review the performance of ICA algorithm in resolving these issues.

#### REFERENCES

- [1] Baumol, W.J., Vinod, H.D. "An inventory theoretic model of freight transport demand", *Management Science*, 16, 1970, 413-421.
- [2] Yokoyama. "Integrated optimization of inventory - distribution systems", *proceedings of eighteenth International conference on computers and industrial engineering*, 10, 1995, 479-484.
- [3] Kim, JU, Kim, Y. "A Decomposition Approach to a Multi-period Vehicle Scheduling problem", *Omega, International Journal of Management*, 27, 1999, 421-430.
- [4] Asgari, a., Aghdasi, d., "a model of integrated inventory and transportation planning in supply chain management: Iran Khodro", the first national conference on logistics and supply chain, Tehran, Iran,2004,10-19.
- [5] Jang, YJ, Jang, SY, Chang, BM, Park, J. "A combined model of network design and production / distribution planning for a supply network", *Computers and Industrial Engineering*, 43, 2012, 263 - 281 .
- [6] Karabuk, S. "Modeling and optimizing transportation decision in a manufacturing supply chain", *Transportation Research part E*, 43 (4), 2007, 321 - 337.
- [7] Chou, A.S.C., Cheng, T.C.E., Lin, B.M.T. "Scheduling in an assembly type production with batch transfer", *Omega*, 35 (2), 2007, 143-151.
- [8] Zegordi, SH, Kamal Abdi, IN, Beheshti Nia, MA "A novel genetic algorithm for solving production and transportation scheduling in a two-stage supply chain", *Computers & Industrial Engineering*, 58, 2010, 373-381.
- [9] David, N., surico, M., Turchiano, B., Kaymak, U. "Genetic algorithms for supply-chain scheduling: A case study in the distribution of ready-mixed concrete", *European Journal of operational research*, 177 2007, 2069-2099.
- [10] Zandieh, Z., Molla Alizadeh, S. "synchronizing production and air transportation scheduling using mathematical programming models", *Journal of computational and Applied mathematics*, 230, 2009, 546 - 558.
- [11] Ghasemi, n., Joiner, M. "multiplexing orders and the timing of mathematical modeling send to help in the supply chain", the National Conference of logistics and supply chain, Tehran, Iran,2011,15-35.
- [12] Chen, A.Z., Liu, S.C. "Variable neighborhood search for the inventory routing and scheduling problem in a supply chain", *Expert system with application*, 39, 2012, 4149-4159.
- [13] Pundoor, G., Chen, Z.L. "Joint cyclic production and delivery scheduling in a two - stage supply chain", *production Economic*, 119, 2013, 55-74.
- [14] Rohit Bhatnagar., Peeyush Mehta., Chee Chong Teo. "Coordination of planning and scheduling decisions in global supply chains with dual supply modes", *Production Economics*, 131, 2011, 473-482.
- [15] Elimam, AA, Dodin, B. "Project scheduling in optimizing integrated supply chain operations", *European Journal of Operational Research*, 224, 2013, 530-541.
- [16] Hishamuddin, H., Sarker, RA, Essam, D. "A recovery model for a two-echelon serial supply chain with consideration of transportation disruption", *Computers & Industrial Engineering*, 64, 2014, 552-561.
- [17] Osman, Hany., Demirli, Kudret. "Economic lot and delivery scheduling problem for multi-stage supply chains", 136, 2012, 275-286.
- [18] Kim, Y.D., Kang J.H. 2009. "Coordination of inventory and transportation managements
- [19] Zegordi, SH, , Beheshti Nia, MA " Integrating production and transportation scheduling in a two-stage supply chain considering order assignment ", *The International Journal of Advanced Manufacturing*, 44, 2009, 928-939.

# Wiki-Based Stochastic Programming and Statistical Modeling System for the Cloud

Vaidas Giedrimas

Department of Informatics  
Siauliai University  
Siauliai, Lithuania

Leonidas Sakalauskas

Department of Informatics  
Siauliai University  
Siauliai, Lithuania

Marius Neimantas

Department of Informatics  
Siauliai University  
Siauliai, Lithuania

Kestutis Žilinskas

Department of Informatics  
Siauliai University  
Siauliai, Lithuania

Nerijus Barauskas

Imagine Communications Group  
Dublin, Ireland

Remigijus Valčiukas

Department of Informatics  
Siauliai University  
Siauliai, Lithuania

**Abstract**—Scientific software is a special type of software because its quality has a huge impact on the quality of scientific conclusions and scientific progress. However, it is hard to ensure required quality of the software because of the misunderstandings between the scientists and the software engineers. In this paper, we present a system for improving the quality of scientific software using elements of wikinomics and cloud computing and its implementation details. The system enables scientists to collaborate and make direct evolution of the models, algorithms, and programs. WikiSPSM expands the limits of mathematical software.

**Keywords**—Wikinomics; open source; mathematical programming; software modeling; online computing

## I. INTRODUCTION

Scientific software is a special type of software because its quality depends not only on the financial results but also that of the quality of appropriate scientific conclusions and the speed of scientific progress. However, the success ratio for the projects of scientific software development is close to average. This means that part of the projects fail, over budget or give inadequate product.

The misunderstandings between final users (scientists) and contractors (software engineers) are even more frequent as usual. This is because of the fact that software engineers sometimes are unable to get deep knowledge of user's domain (e.g. high energy physics or life sciences). In order to avoid possible problems, scientists sometimes try to develop software indecently. However, such projects can also fail due to the lack of knowledge of software engineering domain. For example, scientists may not know (or just do not care about) good software engineering practices, common processes, etc. They may even lack knowledge about good practices or good artifacts of the software, made by their colleagues.

We believe that this problem must and can be solved using the idea of *Wikinomics*, which is introduced by Tapscott and Williams [7]. Wikinomics (or Wiki economics) is a spatial activity, which helps to achieve results using the available resources only. The idea of Wiki-based systems is very simple: the project leaders collect critical mass of volunteer who are willing and who can to contribute small parts. Sum of such small parts makes a huge contribution to project goals and makes this form of mass-collaboration very attractive. Such systems as Wikipedia or Wikitravel are good advocates of the advantages of the Wiki technologies.

On the other hand mass-collaboration in the software modelling or software coding states is not enough. In order to use all (or at least part of) computational power of the distributed infrastructures, we need software developing solutions, oriented to clouds and grids.

Software synthesis for grid, cloud, and other distributed systems is one of the three main distributed computing-related research areas of Siauliai University [1, 2]. On the umbrella of this research area, a new project at Siauliai University was started from October 1, 2010. The main goal of this project was to develop the environment for scientific software synthesis using grid, cloud, web, and wiki-oriented technologies.

The goal of this paper is to present results of the empirical research related to this project, namely - the system for the Stochastic Programming and Statistical Modelling System empowering the members of scientific community with the abilities to make, edit, and reuse models, algorithms and the software.

The rest of the paper is organized as follows: Section 2 outlines the overall architecture of our system including its main components. Section 3 presents the result of testing.

Section 4 describes the comparison of similar systems. Finally, the conclusions are made and future work is discussed.

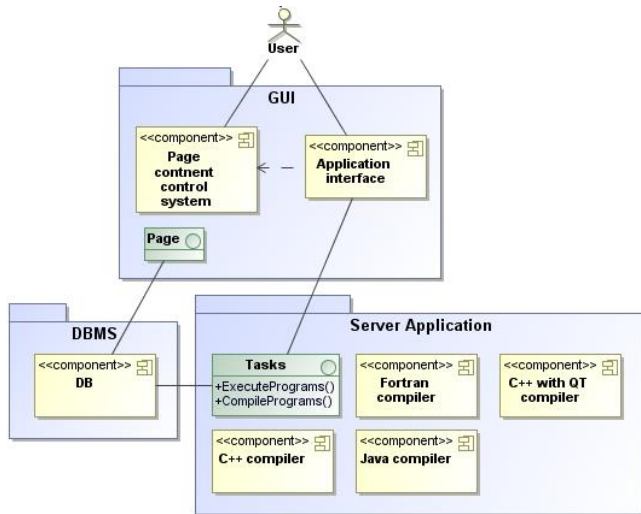


Fig. 1. Main components of the Wiki-based Stochastic Programming and Statistical Modelling System

## II. THE ARCHITECTURE OF STOCHASTIC PROGRAMMING AND MODELLING SYSTEM

We have started from the hypothesis that using together wiki-based technologies, software synthesis methods, and the power of the grid/cloud infrastructure, scientific software can be developed more rapidly and the quality of the software will increase. The project consists of three main stages:

1) *The development of the portal for the wiki-based mass-collaboration. This portal will be used as the UI enabling scientists to specify the problems for software development, to rewrite/refine the specifications and software artefacts given by other researchers, to contribute all software developing for particular domain process. As the target domain for software development we have chosen the set of the statistical simulation and optimization problems. In the future the created environment can be applied to other domains.*

2) *The development of the model of the interoperability bridge between wiki-based portal and the Lithuanian National Grid Infrastructure or other distributed infrastructures. For this purpose currently the private cloud is created at Siauliai University, based on Ubuntu One.*

3) *To refine existing methods for software synthesis using the power of distributed computing infrastructures.*

This paper mainly covers the results of the first two stages.

The system for Stochastic Programming and Statistical Modelling based on Wiki technologies (WikiSPSM) is built using the Model-View-Control architecture. It consists of the following parts (Figure 1):

- Web portal with the content management system and Wiki-capabilities as the Graphical User Interface (see Subsection B for more details).
- Server application for jobs processing (Subsection C).

- Database and its management system for the storage and processing of existing software artifacts including programs, subroutines, and models (Subsection A).

The importance of each part and its role in overall functionality of the system is detailed in the remaining subsections.

### A. Database

The database is based on MySQL. It consists of 34 tables, which can be divided to 5 groups: user data, logging data, job data, website data, and program data.

The tables of the user data consist of information about user access rights, groups, etc. Tables for logging data consist of the portal usage history data, portal user actions, and so on. Job data tables carry information about the submitted jobs and their states and location details.

Website data is the basis for CMS and user interface of the Wiki part of the portal. Program data (PD) is probably the most important (and most distinctive) part of all the systems. PD tables are for storage and processing of the programs, written by users and APIs. In other words, PD is the basis for the Wiki part of the portal.

### B. Wiki-based Portal

The user interface portal is the central entry-point for the users in our system. It consists of four main components:

- *Template-based generator of web pages.* This component empowers a user to make web page content using its structure from the template. The same component is used for the storage of generated web-pages including version control.
- *WYSIWYG text editor.* The functionality of this component is beyond simple text editor on the web page. The component is dedicated to describing mathematical models and numerical algorithms. For safety reasons, it is enriched with the text pre-processing algorithms, which prevents code injection and hijacking attacks.
- *Integrated developing environment* component is implemented for the modeling and the coding of the software. It is built according to theoretical background published in [10].
- *Mathematical Functions Repository.* This component enables a user to retrieve, rewrite, and add repository of mathematical functions with new artifacts, for example, new APIs. Almost all WikiSPSM is built from scratch. However, for efforts optimization it was decided to include “standart” library of mathematical functions in it. Our system is based on NetLib repository LAPACK API. However, it can be changed on demand to work with other libraries, for example, ESSL (Engineering Scientific Subroutine Library) or Parallel ESSL [6].

The architectural decision to store all the mathematical models, algorithms, programs, and APIs in central database makes WikiSPSM easy extensible and evolvable.

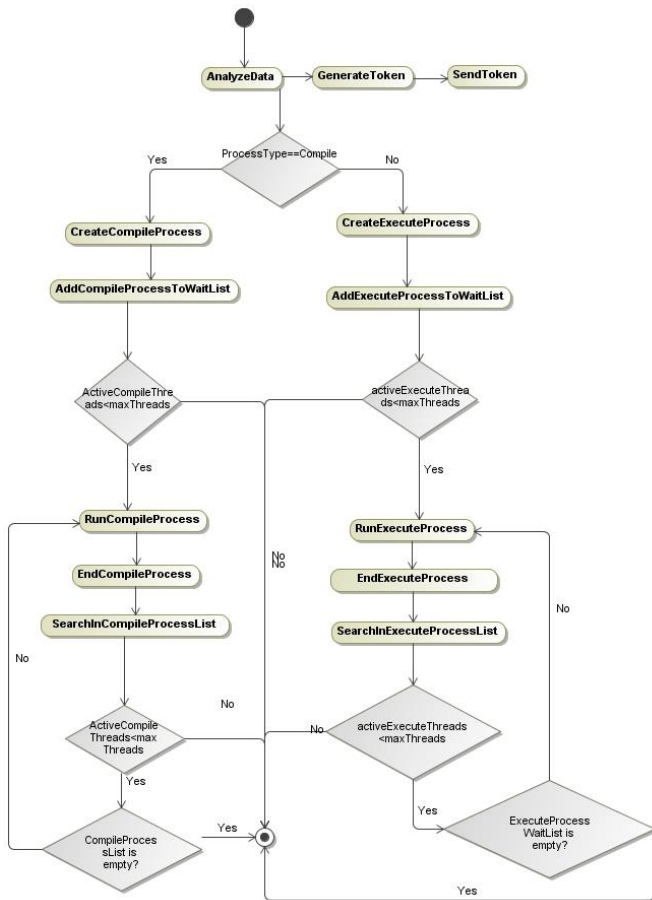


Fig. 2. The algorithm of data analysis and job processing

### C. Software Generation Process

The architectural decision to enable users to write their software in C/C++, Java, Fortran 90, and QT programming languages was made in the design stage of WikiSPSM. In order to implement this idea, command-line interfaces were chosen as the architecture of communication between the UI and software generator part. Software generator must perform the following actions:

- Compilation,
- Job submission (to cloud or to single server),
- Job monitoring,
- Job and its results control.

External command-line compilers were chosen for the compilation of the programs. Theoretically, our system can be configured to work with any programming language if only its command-line compiler (or compilation service) is available. The requirement to have command-line user interface is necessary for user's programs also. In the current version, WikiSPSM is not able to work with the external programs using GUI.

It is planned that compilation and execution must be made on the server side. Note that "server side" here means not only one server, but can be considered as a cloud also (see

Subsection D). The server application consists of three classes: Server, Task, and vrMath. At the application initialization, time object vrMath is created, and then it creates the object Server. These objects help to monitor jobs and to manage them interactively, including abort operation.

The algorithm of the program data analysis and job processing is shown in Figure 2. Object Server creates an object Task for each submitted data array received from the portal. Task object parses the data and sends a Token back to the user (via portal). Token is necessary to identify the job, monitor it, and get its results. When the process Task is finished, it passes the data to Server object and then the process of compilation of execution begins. All the jobs are scheduled and queued. If number of resources (e.g. number of CPU) is less than requested, job waits the end of some other processes. After job completion, its result is stored in the database.

### D. Bridge to Distributed Systems

In its early stage, WikiSPSM was implemented as a server-based distributed program [1]. However, it was observed that increased number of users and submitted tasks have negative impact on the performance of the system. For example, while testing the instability of the system, it was observed when a job tried to use more than 3 CPUs on one server. In order to solve this issue, the architecture of the system has been changed.

In revised architecture of the system. Wiki-based portal and other components remain more or less the same. However, the software generation component is changed dramatically. The adoption of this legacy component is made in two stages:

- *Transformation between different operating systems.* Initially, the server-side application was hard coupled with the Windows operating system, because of used before command-line compilers and Qt library. All this part was redesigned and new Linux-based implementation is made. The changes not only improved the system but also made it completely multiplatform.
- *Transformation between the paradigms.* For better throughput of computing application, server was redesigned to schedule jobs in wide-distributed environments, namely Ubuntu One and Open Stack private clouds (Figure 3). Besides the central DB distributed file system with NFS file system is used for the communication of working nodes. During the development and for the testing a set of VirtualBox virtual machines have been used.

Early test of redesigned component shows very good results (see Section 3).

## III. TEST RESULTS

WikiSPSM is tested both in technical and scientific levels. The set of programs was made using Fortran 90, Java, C, C++, and NetLib repository LAPACK API.

The set of scientific tests consisted of a) the comparison of the results of well-known problems with the results given by the system; b) the comparison of the results of sequential and parallel implementations of the same Monte Carlo method-



based algorithm. All the tasks finished and show expected results only. The participants of the pilot project named the possibility to change and extend programs from the repository as a very useful and efforts-sparing option.

The set of technical tests consisted of the tests of program compilation, execution, data exchange, and user-computer interface support processes. For the system load testing, a program for Monte Carlo algorithm implementation was chosen. On its basis, 500 requests for the server side were generated. The server with 4 CPUs was chosen as an abstraction of the distributed system. The results of this test show (Table 1) that the system acts as a common task distribution system and the given job submission and job execution times fit the Amdahl's law very well.

For 150 jobs, Monte-Carlo problem using new (bridged to distributed systems) execution component was solved two times faster than initial server-based application component. The "toy example" (calculation of the factorial of big numbers) – was solved eight times faster.

#### IV. RELATED WORK

Our system (in short WikiSPSM) has been compared to other commercial and open-source products. Here the comparison with two most distinguished products (each present the set of the tools) is presented.

##### A. Mathematica

Wolfram Research Mathematica is a multiplatform tool with high functionality. It consists of the core and user interface components. The core is able to parse Mathematica programming language, to call external command-line software, process with files using default programs, commute via e-mail, and perform dynamic linking of the functions, in such a way as to extend the functionality of the core. The core with UI communicates using MathLink protocol. For the communication with external software .NET/Link and Java/Link are offered, which enables to call Mathematica's functions from e.g. Java program and vice versa [3, 8].

TABLE I. THE RESULTS OF SYSTEM LOAD TEST

Number of processes	Time for all job submission	Time for all jobs completion
1	12 s	14 min. 21 s
2	16 s	9 min. 41 s
3	22 s	9 min. 51 s
4	NA	NA

TABLE II. THE CRITERIA OF COMPARISON

Criteria	Weight coefficient
Reusability of Mathematic functions, models and programs	0.3
Possibility to extend system repository by new functions, models and programs	0.2
User interface	0.2
Programming languages support	0.1
Programs' execution	0.1
The possibility to integrate system in other systems	0.1

There are two tools under Mathematica umbrella, which can be used as components for a stochastic programming and Statistical Modelling System: *Grid Mathematica* and *Web Mathematica*.

Grid Mathematica is able to perform calculations using distributed infrastructures. This tool is the subject of additional licensing and supports up to 16 cores per task.

Web Mathematica is based on Java applets and servlets. It can perform asynchronous computations only and is useful as a different type of UI only.

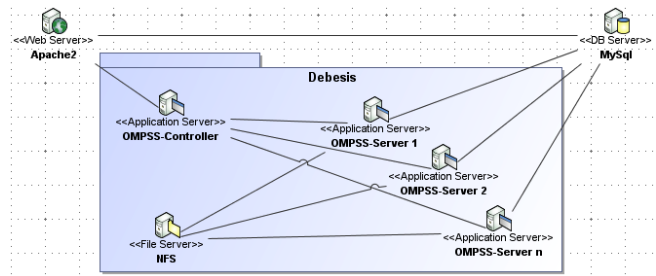


Fig. 3. The architecture of WikiSPSM with the cloud computing component

Grid Mathematica and Web Mathematica cover only two aspects of our System: distributed computation and web-based interface. However, the users of Mathematica aren't able to make new module dynamically, to add functionality to Mathematica Core dynamically, to reuse programs of other users. In Web Mathematica case, only the site administrator has rights do refining of the core. Therefore the use of Mathematica in Wikinomics context is problematical.

##### B. Scilab

Scilab Scientific is chosen as an example of open-source tool because of the functionality. The tool is oriented to solve problems of linear algebra, matrixes, polynomials, statistics, differential equations etc. Because of this, by the functionality Scilab can be named as a good competitor of commercial systems. Scilab is multiplatform tool; it supports all recent operating systems. Scilab provides links between it and C, C++ and Java. Scilab native programming language subroutines can be called for C, C++, and Java programs. There is a possibility to import Matlab code, because of the similarity of Scilab native language and Matlab language. However, in contrast to Mathematica, Scilab does not have task parallelization component or web UI component [4, 5].

##### C. Comparison

Our system and the two competitors described above have been compared using the list of criteria (see Table 2). Most important criterion is the reusability of artifacts, because on it depend the complexity of scientific software and the amount of time required to make new solutions. Second criterion is the possibility to extend the system with new artifacts: functions, models, and programs. The possibility of using the system artifacts from external sources (programs) is no less important.



These three criteria together have direct relation with the idea of wikinomics in scientific software domain.

All three compared products (see Table 3) have rich set of supported mathematical functions and API's. Mathematica has the most rich list of functions and algorithms for the problems of mathematical programming. WikiSPSM uses NetLib repository LAPACK [9] for C++ and FORTRAN and as a consequence provides more functionality as Scilab.

Users of Mathematica or Scilab can reuse their functions directly. In contrast, WikiSPSM does not offer such possibility, because it is web based and all the programs are executed on the server side, not locally.

As could be predicted, WikiSPSM shows best result on second criterion – the possibility to extend system repository. All artifacts are stored in public database, can be reviewed, reused, and appended. Other systems have different, single user oriented architecture. Moreover, they have only a little possibility to change system functions or extend the core of the system by user subroutines.

Comparing the systems by the user interface criterion it's easy to see that WikiSPSM has UI with a reach not as large as Mathematica or Scilab in terms of common functions. (e.g. possibility to show 2D and 3D graphical objects). However, WikiSPSM in its UI has a lot of Wiki-oriented functions (in contrast to other systems) and web-orientation (in contrast to Scilab).

TABLE III. THE ASSESSMENT OF THE SYSTEMS

Criteria	M	S	W
Reusability of Mathematic functions, models and programs	10	7	8
Possibility to extend system repository by new functions, models and programs	3	3	9
User interface	9	5	7
Programming languages support	6	6	8
Programs' execution	9	5	7
The possibility to integrate system in other systems	10	6	2
Total	7.9	5.4	7.3

M – Mathematica, S – Scilab, W – WikiSPSM (our system). The total value is calculated as a sum of values multiplied by the weight coefficients (Table 2)

By the criterion of programming languages, WikiSPSM is clear winner also because it supports C, C++, Java, QT, and FORTRAN in contrast to others, which support only two to three external languages.

By the last criterion (the possibility to integrate the system itself with other IS), WikiSPSM can be assessed by the lower mark than Mathematica and Scilab. Its integration process and possibility of interoperability is the same as of some other application server-based systems. However, this is not a big disadvantage because our goal was to build this system as a portal, as a central entry-point, not as a component.

## V. CONCLUSIONS AND FUTURE WORK

The following conclusions can be made:

1) *The system of Wiki-based Stochastic Programming and Statistical Modelling is developed and it is adopted to cloud*

*computing infrastructure. WikiSPSM can solve various optimization problems using C, C++, Java, Fortran, programming languages and Qt and Netlib Repository LAPACK libraries. The system already used to solve two stage stochastic programming problems.*

2) *The system enable scientist to collaborate and make direct evolution of the models, algorithms and programs. WikiSPSM expands the limits of common mathematical software.*

3) *WikiSPSM can be considered as the job submission and management system for private academic cloud or other distributed infrastructures. It has all the necessary components despite the fact that the system is build from scratch using open source technologies only.*

The results of the project will have direct positive impact on development of scientific software. Since the gap between the two technologies is bridged, each of them promises good performance. The power of the wiki-technologies will ensure the ability of the interactive collaboration on software developing using the terms of particular domain.

Despite the success of the project, some future work is still required:

1) To increase the level of the security, automatically inspect what files and data are stored (or going to execute) in the server side and prevent hazardous actions.

2) To facilitate the configuration of the system and in such way to expand the list of supported programing languages.

3) To refine job scheduling, distribution and parallel execution algorithms by adding stochastic elements.

4) To make a bridge between WikiSPSM and European Grid Infrastructure or commercial infrastructures.

Continued testing of WikiSPSM is planned by solving the following problems: power plant investment planning by stochastic programming; two-stage stochastic programing problem for short term financial planning.

## REFERENCES

- [1] V. Giedrimas, L. Sakalauskas, K. Žilinskas. Towards the environment for mass-collaboration for software synthesis, 2011. EGI User Forum 2011. pp. 1-2 [online] URL: <https://indico.egi.eu/indico/event/207/session/15/contribution/117/material/slides/0.pdf>
- [2] V. Giedrimas, A. Varoneckas, A. Juozapavicius. "The grid and cloud computing facilities in lithuania." In *Scalable Computing: Practice and Experience*. vol. 12(4). pp. 417–421, 2011.
- [3] S. Steinhaus. Comparison of mathematical programs for data analysis. Munich, 2008.
- [4] M. Baudin. Introduction to Scilab. The Scilab Consortium, 2010.
- [5] C. Bunks, J.-P. Chancelier, F. Delebecque. C. Gomez, M. Goursat, R. Nikoukhan and S. Steer. Engineering and scientific computing with Scilab. Birkhauser Boston, 1999.
- [6] ESSL and Parallel ESSL library [online] URL: <http://publib.boulder.ibm.com/infocenter/clresctr/vxxr/index.jsp>.
- [7] D. Tapscott, A. D. Williams. Wikinomics – How mass collaboration changes everything. Atlantic Books, 2011.
- [8] Mathematica Documentation Center [online] URL: <http://reference.wolfram.com/mathematica/guide/Mathematica.html>.

- [9] V.A. Barker et al. LAPACK User's Guide: Software, Environments and Tools. Society for Industrial and Applied Mathematics. 2001
- [10] L. Sakalauskas. Application of the Monte-Carlo method to nonlinear stochastic optimization with linear constraints. Informatica, Vol 15, No 2, pp. 271-282, 2004.

AUTHOR PROFILE

**Vaidas Giedrimas** is a member of the IEEE and the IEEE Cloud computing SIG, a member of Lithuanian Computer Society. He holds PhD from Vytautas Magnus University (Lithuania) in Informatics. He is associate professor of Department of Software Systems at Siauliai University, Lithuania. His research interests include Distributed software systems (grid and cloud computing, component-based software engineering, service-oriented architecture) and methodologies of automated software development. He is the member of the Management committee of COST Action IC1201 Behavioural Types For Reliable Large-Scale Software Systems (BETTY) and Lithuanian representative at NorduGrid and Baltic-HPC associations.

**Kestutis Zilinksas** is a member of the Council of the Lithuanian Operational Research Society. He holds PhD from Vytautas Magnus University (Lithuania) in Informatics. He is associate professor of Department of Informatics at Siauliai University, Lithuania. His research interests include stochastic programming, Monte Carlo method, and numerical methods in optimization.

**Leonidas Sakalauskas** is the main author of the presented project idea. He holds the professor position at Department of informatics, Siauliai University and principal researcher position at Operational research sector at Systems analysis department. He also is the president of the council of the lithuanian operational research society. His research interests consist of operational research, artificial intelligence, data mining and information systems (business informatics).

**Nerijus Barauskas** holds MS from Siauliai University (Lithuania) in Informatics. He has rich experience as Web Application Developer and Cloud computing infrastructures builder in Lithuania and Ireland. Currently he is working as System Developer at Imagine Communications Group (Ireland)

**Marius Neimantas** holds MS from Siauliai University (Lithuania) in Informatics. Currently he is working as Chief specialist at Informatics department, National Law Administration (Lithuania). Security is the main focus of his research interests.

**Remigijus Valčiukas** holds MS from Siauliai University (Lithuania) in Informatics. Currently he is working as Software Developer at Šiaulių bank (Lithuania). His research interests consist of software development with SQL, C++, QT, PHP, JavaScript and jQuery programming languages, Oracle and other databases.

# Multi Agent Architecture for Search Engine

Disha Verma  
Research Scholar,  
Jagannath University,  
Jaipur, India

Dr. Barjesh Kochar  
Professor (IT),  
Jagan Institute of Management Studies  
Delhi, India

**Abstract**—The process of retrieving information is becoming ambiguous day by day due to huge collection of documents present on web. A single keyword produces millions of results related to given query but these results are not up to user expectations. The search results produced from traditional text search engines may be relevant or irrelevant. The underlying reason is Web documents are HTML documents that do not contain semantic descriptors and annotations.

The paper proposes multi agent architecture to produce fewer but personalized results. The purpose of the research is to provide platform for domain specific personalized search. Personalized search allows delivering web pages in accordance with user's interest and domain. The proposed architecture uses client side as well server side personalization to provide user with personalized fever but more accurate results. Multi agent search engine architecture uses the concept of semantic descriptors for acquiring knowledge about given domain and leading to personalized search results. Semantic descriptors are represented as network graph that holds relationship between given problem in form of hierarchy. This hierarchical classification is termed as Taxonomy.

**Keywords**—Search engine; Data mining; Multi agent systems (MAS); Semantic mapping; Hozo

## I. INTRODUCTION

Internet is one of the biggest repositories of information of web documents or HTML documents. Various search engines were created to access these web pages on basis of keywords entered by user. The main drawback of these traditional search engines is that they perform searching by focusing on keywords rather than on meaning of content. It is the reason that they lack semantic dependency among user entered keywords.

The users are getting addicted to latest technology trends and they want their services at their doorstep. Due to diverse search results, the accuracy of retrieved information is decreasing. Different users have different interests but owing to algorithm without interest as a parameter search engines produce same result for all users that does not satisfy individual user needs. This led to development of personalized and semantic search engines. With the advent of Vertical search engines also called Vortals, the personalized search engines were the next palpable step. Vertical search engines (VSE's) emphasize on single topic or specific domain while personalized search engines extracts user preferences and provide them tailored results.

The search engine proposed is an intelligent domain specific personalized search engine which personalizes the results pertaining to a specific domain. The concept which

differentiates this research from previous works is personalization at both ends (client and server) and domain specific searching. Unlike previous engines the proposed architecture keeps the sophisticated architecture of Google and then adds layers of personalization to it. In the process of adding layer it doesn't let to significant increase of processing time. Thus making the engine more personalized results in reasonable time.

The first part of the paper represents previous work in the field followed by theoretical explanation of the concept introduced for the architecture (Evolution of Ontology in Multi Agent Systems). Problem definition section explains the current problem and proposed solutions for the area. Following the problem definition section is proposed architecture section which discusses the proposed architecture in detail. After which the analysis and evaluation section compares the traditional search engine with the proposed architecture. The last section explains the future scope and concludes the research.

## II. LITERATURE

Various studies have been conducted and results are deduced regarding personalized view of search engines as follows:

Gauch, & Chaffee [1] captured user interest by browsing their history stored in form of cookies. Speretta & Gauch [2] developed personalization by identifying user profiles. The personalization model used methods to grab user's search history in order to retrieve results precisely. Chirata & Nejidi [3] found that queries are handled in three categories viz. clear queries, semi-ambiguous and ambiguous queries. There is no need of personalized search in handling non ambiguous queries but for handling semi ambiguous and ambiguous queries, personalization is necessary. According to [19], the user based personalization is the most significant contextual factor, which can alleviate an ambiguous web search in an ad hoc retrieval task. Some works model the user interests as being a rich repository of personal information extracted from the user search history like past click through data [20], browsing features or desktop information [21], etc. The use of multi agent technology is widespread among large number of applications like distributed systems, data mining, logistic management and sensors [4]. The technology is based on working of nature of agents to be employed for retrieval of information. Each agent has its independent tasks and they work in collaboration with other agents in system. Another variant of multi agent systems is non-equilibrium systems that have been studied by Nicolis & Prigogine [5]. Non-

equilibrium system means previous keywords and phrases related to query can be modified during next phase of searching. [14] investigated, compared various Text categorization algorithms and proposed profile based personalization using Open directory project. [15] proposed the search engine which took in account positive as well as negative preferences of the user. [16] proposed click content and location entropies to measure interest in the content and/or location information in the results. [18] proposed a special log system to capture users' interactions in web systems. The system consisted of two subsystems: log capturing and log analyzer system. The log system makes use of several attributes identifying user session, such as embedded session identifiers, cookies and IP-address of the accessing host. All the above researches were based on HTML pages. Finding semantic categorization was difficult. [17] proposed an architecture based on creating profiles using semantic descriptors. This paper proposes a similar architecture.

### III. EVOLUTION OF ONTOLOGY IN MULTI AGENT SYSTEMS

With advent of concept of Ontology and Semantic Web, Knowledge Management (KM) solutions have started employing in different environments where agents can define their own ontology. Ontology allows an agent to perform following tasks:

- Agents can distribute collection of data and enable communication of data in different environments and applications.
- Agents increase information retrieval performance by searching results in less time.
- Agents use learning algorithms to give examples regarding retrieved results on basis of query entered by user.
- Interface Description Languages and services are provided for different environments where Interface Language refers to defining of data objects and their location.
- Agents are able to access some distributed models to enable interaction between processes like CORBA (Common Object Request Broker Architecture), RMI (Remote Method Invocation).

#### A. Multi Agent Systems (MAS)

1) *Problem:* - While using KM solutions for integration of knowledge processes, there used to work on Peer to Peer (P2P) architecture for enabling distributed control of knowledge resources at centralized location via peers. Peers are directly connected to Hub. But this architecture is not suitable for accessing knowledge from complex systems.

2) *Solution:* - Multi Agent Systems (MAS)

3) *Analysis:* - Systems where individual self-authorized agents derive new facts with the help of other agents are called MULTI AGENT SYSTEMS.

In these systems, individual agents create their different models and prototypes instead of following standard ontology. An agent is defined as an entity with/without body. Agents use

different kinds of knowledge sources and resolve differences among themselves to provide best answers of queries in complex KM environment.

Ontology is evolved by learning concepts and relationships by taking guidance from other agents. Agents are trained explicitly as well as they learn in the course of their use. With time agents improve their knowledge also communication among various agents adds to their learning. Fig.1. represents the characteristics which are imperative for software agents.

4) *Approach involved:* - There are following assumptions to be kept in mind which are as follows:

- Assume given organization has  $n$  agents  $Ag_1, \dots, Ag_n$  where each agent manages knowledge for its organization.
- Each agent knows concept which is denoted by  $C_k$ .
- Each agent has positive and negative thoughts ( $ti$ ) with respect to that concept.
- Each agent has learning algorithm ( $Li$ ) or classification mechanism.
- Each agent has its unique Ontology denoted as  $O_i$ .
- Each agent has set of features for representing concepts. It is denoted as  $fi$ .

System is represented as:  $Ag_n = \{Li, ti, fi, O_i\}$

a) *Learning Algorithm (Li):* - An agent learns a concept under supervision of teachers. It is also possible that teachers are not well expert about given queries, in that case they can use learning algorithm to give example regarding queries.

b) *Set of thoughts/examples (ti):* - In this, positive and negative examples are taught to agents by teacher. These examples are related to given problem domain. Using this classification capability, learner agent is able to decide whether example is positive or negative.

c) *Set of features (fi):* - They are most important factor to represent concepts. Features are selectively used by each agent to represent different concepts. In Multi Agent concept learning, we have to collect most related features from different sources of knowledge in order to develop a comprehensive concept.

d) *Ontology (Oi):* - In terms of concepts and knowledge sharing, it is defined as mixture of Meta concepts and fine grained structure. Meta concepts are concepts which are divided into sub concepts until fine grained concept level is reached. It is preferred to use own ontology rather than using standard ontology.

Ontology- based semantic integration [6,7, 8] can be achieved to resolve differences that arises during run time interaction of system and agents. It is implemented in following ways:

- Using a single centralized ontology for all agents and application domain.

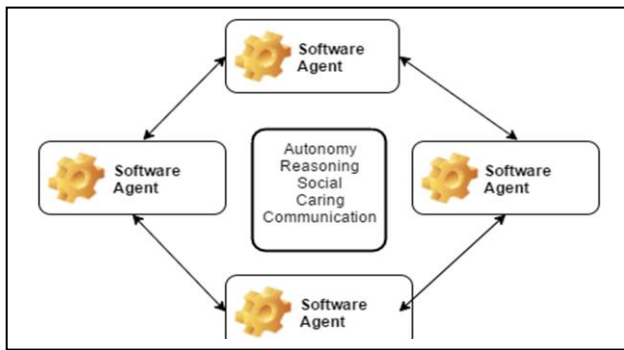


Fig. 1. Characteristics of the agents

- Merge source ontology into a common ontology to prevent overlapping of concepts by various agents.
- Search set of mappings or matches between two ontology when it is difficult to merge them due to the difference.

### B. Semantic Mapping on Ontology

1) *Importance:* - Multi agent systems are used to convert input query into semantic annotations and computes their rank by semantic mapping. Then, there is need of mapping in ontology because Ontology gives complete description of a given problem that can be communicated among people and application systems. Each concept has its own attributes and relationships.

2) *Semantic Mapping:* - It is defined as process of finding solution to given problem with the help of relative concepts used in other domain. It is needed in order to realize full growth of Semantic Web and processes information between ontology. Consider a scenario: - There are two beers belonging to different countries. One is Australia and other is Denmark.

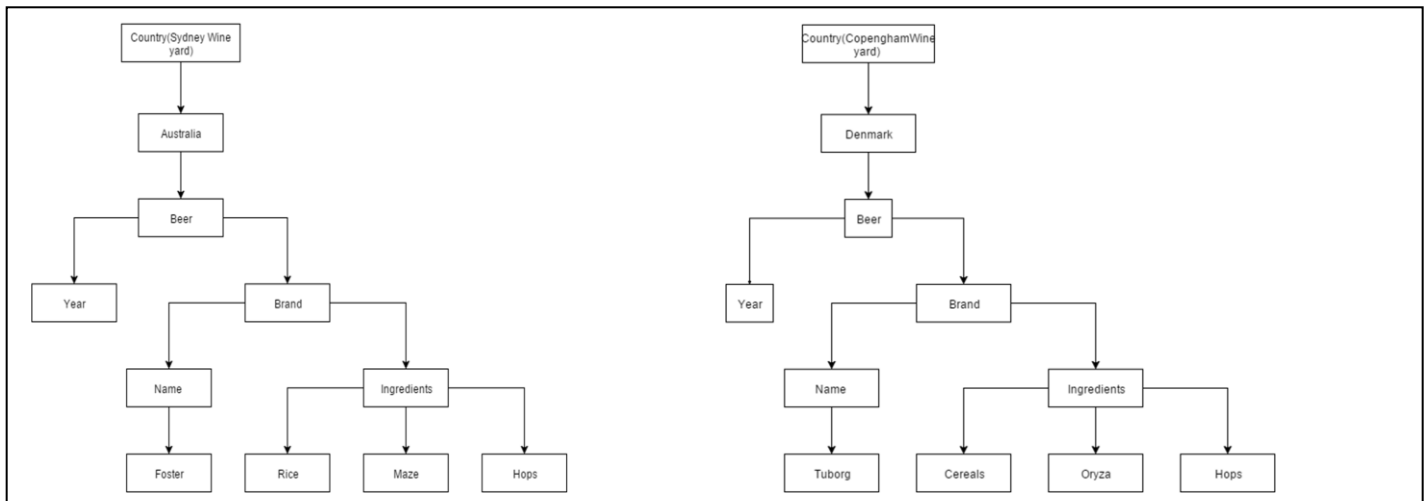


Fig. 2. Semantic Mapping[9]

Let two countries develop their web pages and decided to enable their web content. They make use of ontology in making web pages. Here is mapping between the ingredients of two beers belonging to different countries. Both beers contain different ingredients but some are common to both due to which we can find suitable beer and give validation to concept of Semantic Mapping. Figure 2 represents semantic mapping of sample beer. We have seen that Rice and Oryza are treated as equivalent concepts in different countries. Oryza is biological name of rice. Such a correspondence is called Semantic Mapping. Thus because the context was same the semantic matching could be achieved. The context was interpreted using ontology.

### IV. PROBLEM DEFINITION

Most of search engines are based on client server approach. The personalization is implemented either at client side using histories, location or click model [16]. This approach suffers from the fact if the machine is used by multiple users or user chooses to delete the history. The other approach focuses on creating profiles explicitly for each

individual user and storing these profiles on server. This approach requires a sophisticated system of storing and fetching profile information in small amount of time. Also this approach raises a concern if user is not interested in explicit profile. There is a need to create an optimum level of dependency on both client and server side to achieve favorable personalization.

The next area of concern is efficiency. Since our architecture works on fetching results from basic search engine like Google and then applying personalization the added layer will increase the response time. The higher the response time lower the efficiency. So it is essential to introduce a technology which reduces this time to achieve personalized yet quick results.

### A. Plan of Solution

- **Introducing Levels of Personalization:** The solution of first problem is achieved in our paper by introducing four levels of personalization:

Level 3: Personalization based on both explicit profile and implicit profile.

Level 2: Personalization based on only explicit profile (history disabled)

Level 1: Personalization based on only Implicit profile (based on history )

Level 0: No personalization.

Level of personalization decreases from Level 3 to 0. Depending upon user requirement the personalization can be achieved using both( client side personalization & server side personalization), either or none.

- The technology proposed for second problem is in-memory data grid or distributed cache. An in-memory data grid (IMDG) is a data structure that resides entirely in RAM (random access memory), obviating the need for electromechanical mass storage media. The memory is divided into redundant nodes. The data is processed using these nodes. Initially complex problem is divided into several small subparts and then executed simultaneously. This not only decreases the processing time significantly but also makes it practical to store terabytes of data completely in RAM.
- Another significant role in the architecture is played by Agents. The use of multiple agents refines the results which are then passed to semantic descriptor module that will parse entire results and convert them into semantic blocks of data so that they can be mapped to given ontology. Mapping to ontology requires ontology development phases like specification phase, design and formalization phase. All these techniques are incorporated in proposed multi agent search engine architecture.

## V. PROPOSED APPROACH

The proposed semantic multi agent search engine processes query retrieved from traditional search engines and analyzes these results according to user's priority with the help of multiple agents. This search engine is beneficial because it has ontology domain module which is used to represent relationship between user's preferences and produced search results.

### A. Modules Interaction

The search engine has three modules namely Agent module (consisting of various agents), semantic descriptor module and ontology development module.

*a) Agent Module:* - It involves use of various agents that interacts with user and produces refined results. Agents are interface agent, facilitator agent, resource agent, mining agent and many more. Their functions are listed in table 1.

Pre decision made in order to mine given data sources is called Data Mining Task Planning. Data mining task planning requires compensation between Facilitator Agents and Mining Agents through message passing. Figure 3 represents this interaction

TABLE I. SOFTWARE AGENTS ANF THEIR ROLES

AGENTS	ROLE
(a) Interface Agent (User Agent)	It interacts with user to provide requirements and displays results. It has interface module that contains method for inter agent communication.
(b)Facilitator Agent (Management Agent)	It activates different agents. It receives questions from interface agent and may take the help of group of agents to solve those questions.
(c) Resource Agent (Data Agent)	It maintains Meta data information about data sources. It generates queries based on user request and sends their results to user agent.
(d) Mining Agent	It implements Data mining techniques and algorithms.
(e) Result Agent	It observes mining agents and other results from them. After obtaining results, these agents show results to user agent by integrating with manager agent.
(f) Broker Agent	It is advisor agent that can send reply to query of an agent with name and ontology of respective agent.
(g)Ontology Agent	It maintains and provides knowledge of ontology to solve queries related to ontology

Consider User Agent is denoted U. Facilitator Agent is denoted by X. Broker Agent is denoted by Y. Mining Agent is denoted by Z. If U sends request to X to ask for Data mining with other agents in system. Then X tries to compensate with Y to determine which agents are suitable for performing task. Mining Agent Z is responsible for completion of task while X is used for planning. When Z is completed, it shows results to X and X passes them to U.

*b) Semantic Descriptor Building Block:-* It automatically converts text or sentences into meaningful blocks of data and put them in semantic descriptor which is nothing but graph showing concepts and values. The concepts are derived so that they can be designed to ontology in hierarchical manner.

### *c) Ontology Development Module*

The term Ontology [10] can be defined in different ways. Ontology is abbreviated as FESC (Formal, Explicit, and Specification of Shared Conceptualization) which is defined as:

- Formal: It specifies that it should be machine understandable.
- Explicit: It defines type of constraints used in model.
- Shared: It means that ontology is shared by group. It is not restricted to individuals.
- Conceptualization: It refers model of some phenomenon to identify relevant concepts of that phenomenon.

Ontology is also defined as set of concepts and relationships arranged in hierarchical fashion. Ontology development [11] needs well defined methodology that must follow certain guidelines:

- Ontology being developed should follow Software Engineering standards.



- Ontology development strategy should be simple and practical.

**B. Phases Involved in Developing Ontology**

The phases that are being used in developing ontology also satisfy Software Engineering principles and thus called as Software Development Life Cycle (SDLC) phases. They are described below:

*a) Specification Phase:* - This phase has its few activities.

- Domain Vocabulary definition: - It defines common name and attributes for domain concepts.
- Identifying Resources: - A Resource is anything that has URI. So, if some concepts have number of instances, then they can be grouped into a class.
- Identifying Axioms: - They are structures that represent behavior of concepts.
- Identifying relationships: - Relations are defined within resources.
- Identifying data characteristics: - Defines features of types of resources and their relationship.
- Applying constraints: - Constraints represent named relationships between domain and range class.
- Verification: - After designing preliminary web ontology model, it is necessary that it should be tested for its correctness.

*b) Design Phase:* - The phase is backbone of Semantic Web. The physical structure of designed ontology is based on RDF model which is associated with three triples- Subject, Predicate and Object.

- Predicate: - All characteristics of resources and relationship are taken as Predicate. E.g. each train is assigned unique PNRNo called as ‘HasPnrNo’.
- Subject: - All domain classes of characteristics and relationships of resources are taken as Subject. E.g. there are various passengers travelling to city each having unique URI, so they are grouped in ‘CityPassengersGroup’.
- Objects: - Refers to Range class relationships. E.g. HasPnrNo contains range class ‘NUMBER’ which is literal.

*c) Formalization Phase:* - This phase is result of output of ontology obtained in design phase with the help of some tools.

**VI. EVALUATION AND ANALYSIS**

Firstly, the user entered query related to computer components or appreciations. The search results are produced through GOOGLE search engine. It produces millions of results that are relevant or non relevant.

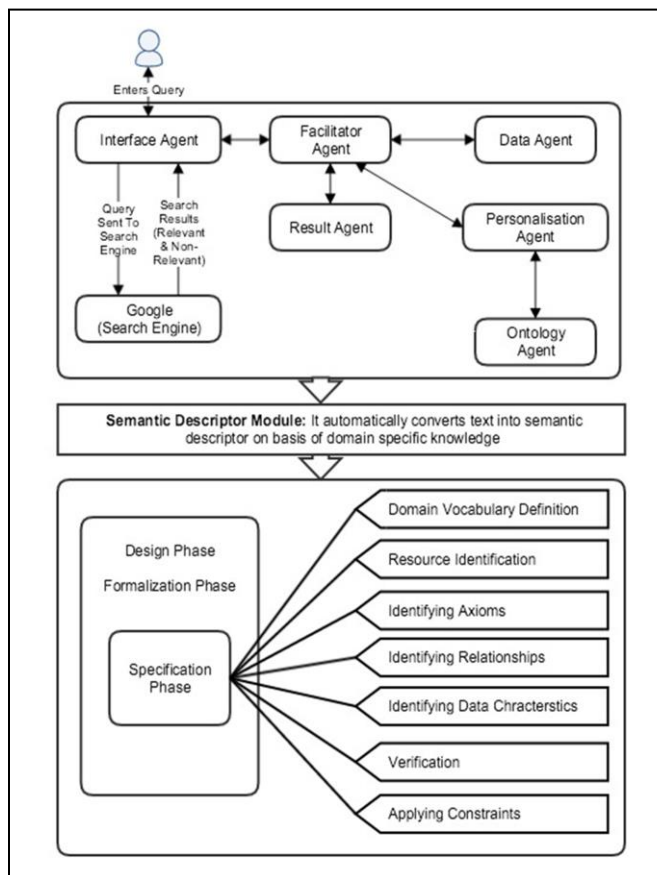


Fig. 3. Outline of proposed Multi-Agent system architecture

Figure 4 represents the complete architecture of the proposed search engine.

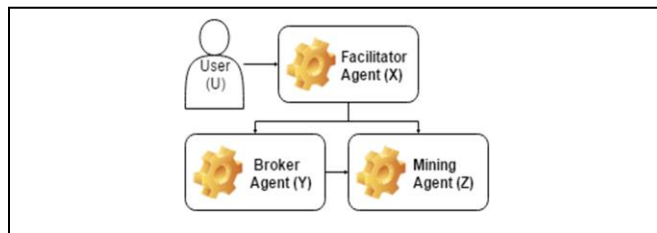


Fig. 4. Flow between Agents

TABLE II. COMPARING EFFECTIVENESS OF SEARCH USING KEYWORDS AND AGENT TECHNOLOGY

Query	Search using keywords (GOOGLE)	Search using agent search engine (semantic descriptors)
Query 1	45%	54%
Query 2	56%	87%
Query 3	64%	85%
Query 4	60%	70%
Query 5	62%	78%
Query 6	44%	65%
Total	55.2%	73.2%

The results are sent to various agents for achieving refined results by involvement of data mining agent as well. Ontology agent is responsible for learning knowledge about given domain. It does not build respective ontology. Then descriptors of search results are built. It can be done automatically. The results are analyzed by block of automatic descriptor building in order to improve accuracy of results. For example, user has searched about computer components. The block of query descriptor building creates the following query descriptor after passing through ontology module. Domain ontology is built on computer appreciations using one of ontology editor called as HOZO. Hozo is different from other ontology editors in following aspects as it is user friendly environment lets users to work easily on it. Hozo has API named as HozoAPIver 1.15 that accesses existing ontology Inheritance information is clear and easily accessible by two options: One is from Super Classes through is-a link. Other is from Class constraint.

#### A. Analysis of search results using GOOGLE and proposed search engine

After manual analysis of results, the first 50 relevant documents results are put in table for both traditional search engine GOOGLE and semantic descriptors used in search engine. The results are given in table II. From table II, it can be said that results retrieved from search engine with semantic descriptors are 20-22% higher than that obtained from GOOGLE.

### VII. CONCLUSION AND FUTURE WORK

A large number of crawlers are present for collecting information from web pages. They retrieve data securely but their distribution and indexing is not satisfactory. The resources are scattered at one central location that creates penalty on network bandwidth. The solution to this bottleneck is use of multiple agents that access information from various pages. The goal of our study in this paper is to propose search engine that allows use of agents to perform various operations in retrieving search results of user. Semantic descriptor is used to map produced results to specific knowledge. The engine is based on ontology domain that is built using ontology development phases to ensure hierarchical representation of results. Use of agent technology allows more refinement in search results thus providing description about concepts used in results. The researched focused on creating personalized result using profiles (implicit and explicit). User is provided with option to choose the level of personalization which provides comfort to the user. Our future work will focus on improving the response time by comparing various technologies. Also we would focus on implementing levels of security to personalization which is still the area of concern for personalization. Users are interested in getting personalized results but owing to security concern of their personal information they are hesitant towards it. So we will try and focus on developing a mechanism which help user to have personalized research keeping the security intact.

#### REFERENCES

[1] Gauch, S. J Chaffee, A ontology based personalized search and browsing. Web intelligence and agent systems, 1 (3), 219-234 (2003).

[2] Speretta M, Gauch S. Personalized search based on user search histories. In Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on 2005 Sep 19 (pp. 622-628). IEEE.

[3] Chirita, P. A, Nejidi, Using odp metadata to personalize search. In proceedings of 28<sup>th</sup> annual international ACM SIGIR conference on research and development in information retrieval, ACM August 2005.

[4] U. Fayyad, R. Uthuruswamy, "Data Mining and Knowledge discovery in databases", *Communications of the ACM*, 39(11), 1996, pp 1-15

[5] Nicolis G., Prigogine I.: Self-organization in non equilibrium systems: New-York, Jonh Wiley & Sons (1977)

[6] Gagandeep Singh, Dr. Mayank Singh, "Implementation of multi agent systems with ontology in data mining", *International Journal Of Research In Computer Application & Management (IJRCM)*, Vol. No. 3, Issue No.1 ISSN 2231-1009", January 2013, pp 111-117

[7] N.F. Noy, M.A. Musen, "Algorithm and tool for automated ontology merging and alignment", *Proceedings 17<sup>th</sup> National Conference on Artificial Intelligence (AAAI-2000)*, Nov 2000

[8] A. Doan, J. Madhavan, A. Halevy, "Ontology matching: machine learning approach", *Handbook on Ontologies in Information Systems*, S. Staab and R. Studer", May 2004, pages 397-416

[9] Gagandeep Singh, Comparative study of search engine and semantic search engine: a survey" in National Conference named NCACT-2013, pp 57-61, organized by Department of Computer Science & Applications, M.D. University, Rohtak (Haryana) on March 30, 2013

[10] Gagandeep Singh, Vishal Jain, "Information retrieval (ir) through semantic web (sw): an overview", In proceedings of CONFLUENCE-The Next Generation Information Technology Summit, 27-28 September 2012, pp 23-27.

[11] Gagandeep Singh Narula, "Ontology development using hozo and semantic analysis for information retrieval in semantic web" in 'CIIP-2013 IEEE Second International Conference on Image Information Processing with IEEE Conference Record Number 31034 on December 9-11, 2013'.

[12] T. Gruber, "The role of common ontology in achieving sharable, reusable knowledge bases" principles of knowledge representation and reasoning: *Proceedings of the Second International Conference, Cambridge, MA*, 1991, pages 601-610

[13] C. Giannella, R. Bhargava and H. Kargupta, "Multi-Agent systems and distributed data mining", *Lecture Notes in Computer Science*, 2005, pages 1-15

[14] Fang Liu, C. Yu and Weiyi Meng, "Personalized Web search for improving retrieval effectiveness," in IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 1, pp. 28-40, Jan. 2004. K. Elissa, "Title of paper if known," unpublished.

[15] K. W. T. Leung and D. L. Lee, "Deriving Concept-Based User Profiles from Search Engine Logs," in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 7, pp. 969-982, July 2010.

[16] K. W. T. Leung, D. L. Lee and W. C. Lee, "Personalized Web search with location preferences," Data Engineering (ICDE), 2010 IEEE 26th International Conference on, Long Beach, CA, 2010, pp. 701-712.

[17] L. Zhuhadar and O. Nasraoui, "Personalized Search Based on a User-Centered Recommender Engine," Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, Toronto, ON, 2010, pp. 200-203.

[18] T. Robal and A. Kalja, "Learning from users for a better and personalized web experience," Technology Management for Emerging Technologies (PICMET), 2012 Proceedings of PICMET '12., Vancouver, BC, 2012, pp. 2179-2188.

[19] T. Park. Toward a theory of user-based relevance: A call for a new paradigm of inquiry. J. Am. Soc. Inf. Sci., 45(3):135-141, 1994.

[20] X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval using implicit feedback. In SIGIR '05: Proceedings of the International ACM SIGIR Conference, pages 43-50, New York, NY, USA, 2005. ACM.

[21] S. Dumais, E. C. J. J. Cadiz, G. Jancke, R. Sarin, and C. R. Daniel. Stuff i've seen: a system for personal information retrieval and re-use. In SIGIR'03: Proceedings of the International ACM SIGIR Conference, pages 72-79. ACM Press, 2003.

# Towards a New Approach to Improve the Classification Accuracy of the Kohonen's Self-Organizing Map During Learning Process

El Khatir HAIMOUDI

Plury-disciplinary Laboratory, Faculty Poly-disciplinary of  
Larache University UAE  
Larache, Morocco

Hanane FAKHOURI

Plury-disciplinary Laboratory, Faculty Poly-disciplinary of  
Larache University UAE  
Larache, Morocco

Loubna CHERRAT

Mathematics and Application Laboratory, Faculty of  
Sciences and Techniques University UAE  
Tangier, Morocco

Mostafa Ezziyani

Mathematics and Application Laboratory, Faculty of  
Sciences and Techniques University UAE  
Tangier, Morocco

**Abstract**—Kohonen self-organization algorithm, known as “topologic maps algorithm”, has been largely used in many applications for classification. However, few theoretical studies have been proposed to improve and optimize the learning process of classification and clustering for dynamic and scalable systems taking into account the evolution of multi-parameter objects. Our objective in this paper is to provide a new approach to improve the accuracy and quality of the classification method based on the basic advantages of the Kohonen self-organization algorithm and on new network functions to pre-eliminate the auto-detected of drawbacks and redundancy.

**Keywords**—Artificial neural networks; self-organization map; Learning algorithm; Classification; Clustering; Principal components Analysis; power iteration

## I. INTRODUCTION

The self-organizing map of Kohonen (SOM), is a model of artificial neural networks (ANN) widely used in different domains, particularly in the classification and clustering of multi-parameter objects. These intelligent systems are characterized by special abilities such as learning, adaptation and the possibility of visualization of Multiparameter objects with a reduced space. Other paradigms can be used, such as retro-propagation network (Counter-propagation) and Back-propagation, both types unlike the SOM map based on supervised learning algorithms that requires advance preparation of desired outputs. Our objective is to reveal the ambiguities and obstacles that may limit the application of this paradigm in the different domains of human activities, and find possible solutions to eliminate them. To reach the desired result, we made a preliminary theoretical study of the learning process, and performed experimental tests. On the basis of analysis of results achieved, we fixed the nature and details of the problem and proposed a solution already practiced and justified theoretically that can eliminate the obstacles detected. At the base of this solution, a new structure of the Kohonen network was realized. The new realization uses a block of preliminary processing of input data. The treatment method is

based on a combination of two well-known algorithms which are used in the field of data analysis. The tests performed and presented show that the proposed model of the map is more advantageous to be used as a means for the creation and the application of intelligent systems in the various fields of human activities.

Except the introduction and conclusion, the work also contains four sections. The introduction presents a resume of this work, which explains the problematic and proposed solutions. In the second section, we give a detailed explanation of the problem of classification and clustering, and present examples of classical and recent methods used to solve this problem. The third section describes the process of classification and clustering using the organizing Kohonen map, also shows the detail of the learning algorithm of this paradigm of the neural network, and fixed observations of its functioning. The improved model of the map is presented in the fourth section. This part explains the detail of our follow up approach to eliminate ambiguities identified in the previous part. Also in this part, we present and explain the changes brought to the standard model map, and the functional structure of the proposed model. The fifth section contains the assessment of this work. A program was developed which is able to present the results of the map's learning on a two-dimensional space, this approach allowed us to properly analyse problems in the standard model and show the advantages of the proposed model concerning the pertinent and accuracy of classification and clustering. The paper concludes by citing the observations fixed in the functioning of the standard model of the map and the advantages of the proposed model and the future prospects.

## II. THE MULTI-PARAMETER OBJECTS CLASSIFICATION AND CLUSTERING

The classification problem is to establish the dependence of the input pattern (a specific object, the phenomenon or process), which is characterized by a vector of parameters (characters), to one of several predefined classes [1].

Examples of these type of problems are handwriting recognition, language recognition, and classification of ECG signals. During clustering, the training sample with class labels is missing and performed detection of the similarity between the patterns, and the similar patterns are brought to the same category. Proximity is often understood in the sense of Euclidean distance [2, 3]. This problem occurs when extracting data, the study of their properties and compression. Thus, two paradigms of classification have been identified: supervised and unsupervised (clustering).

Classical supervised methods are probabilistic methods, in particular, the method based on Bayes formula adapted for manual calculations [4, 5]. Modern methods are not suitable for calculation without using the computers. The application of these methods produces the rules capable to through the given objects to certain specific classes. Rules solutions can be derived as characteristics probability assignment the pattern to a particular class (Bayesian method) or as a simple analytical function (discriminate analysis method). This method has certain limitations, such as lack of flexibility, which is a consequence of the application only of rules of linear solutions [6].

A. *Classic structures and paradigms of learning neural networks for classifying and clustering problems*

New features to classifying and clustering provide artificial neural networks (ANN). Such networks are simplified models of biological neural networks of the brain where neurons are simulated by relatively simple elements of the same type (formal neurons). Neurons are connected together by means of synaptic connections (synapses). The defining features of the ANN are distributed and parallel execution of calculations and learning opportunities.

To solve the problem of clustering ANN is used with unsupervised learning paradigm. Thus, the internal data structure is determined that is the hidden correlation between patterns. The learning process, as in the supervised learning, consists of progressive adjustment of the weighting values. Obviously, such an adjustment may be made only on the basis of information available to the neuron, that is, its status and existing weights. On this basis, and by analogy with the known principles of self-organization of the nerve cells, the learning method of Heb and the method of competitive learning [7, 8] are constructed. It can be assumed that the rules for the classifying and clustering solutions are presented as a set of weights adjusted due to certain iterative procedure of learning.

B. *Competitive learning*

This procedure, which distributes the training set into clusters, is peculiar to the input data [9]. This type of neural network processes only input vector  $X$  and thus implements the learning procedure "unsupervised". An example of neural networks with competitive learning is shown in Figure 1. All neurons in the output layer  $o$  combined with all the neurons of the input layer  $i$  connections with weight coefficients  $w_{io}$ . When the input vector  $x$  is fed to the input of the neural network, only one output is activated ("Winner"). When network correctly learns all the input vectors with the same cluster, there will be one winner. There are two methods to

determine the winner and establish appropriate rules of learning: method of scalar product and the method of the Euclidean distance.

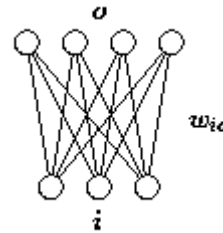


Fig. 1. Simple neural network with competitive learning [45]

**The scalar product method [9]:** Assume that the input vector  $x$  and weight vector  $w_o$  is normalized to unit length. For each neuron in the output layer, the output value  $y_o$  is calculated according to scalar product of the input vector and the vector of weighting coefficients:

$$y_o = \sum_i w_{io} x_i = \mathbf{w}_o^T \mathbf{x} \quad 1$$

In the next step, output neuron  $k$  is elected with the maximum value of the output:

$$\forall o \neq k: y_o \leq y_k \quad 2$$

Then the value of the output neurons of the output layer is set so that  $y_k = 1$  and  $y_{o \neq k} = 0$ . This defines the competitive feature of the network, and the output layer can be called a layer of "winner takes all". The selection function of the output neuron with the maximum output value can be made by the neural network known as Maxnet [10].

When a neuron-winner  $k$  is elected, the weights are corrected in accordance with the formula:

$$\mathbf{w}_k(t+1) = \frac{\mathbf{w}_k(t) + \gamma(\mathbf{x}(t) - \mathbf{w}_k(t))}{\|\mathbf{w}_k(t) + \gamma(\mathbf{x}(t) - \mathbf{w}_k(t))\|} \quad 3$$

Where  $\gamma$ : coefficient of learning (rate learning).

Divider shows that all the weight vectors are normalized. Note that only the weights of the winner neuron  $k$  are changing.

Changing weight coefficients according to formula (3) rotates the weight vector  $\mathbf{W}_o$  in the direction of the input vector  $\mathbf{x}$ . Whenever vector  $\mathbf{x}$  is supplied to the input of the network, the closest weighting vector is elected which subsequently is wrapped in the direction of the input vector. Finally, the weighting coefficient vectors rotate in the direction of areas in space where there are multiple input vectors, that is, clusters of inputs.

**Euclidean distance method [9]:** In the method of scalar product, it is assumed that the input vector  $\mathbf{x}$  and the vector of weighting coefficients  $\mathbf{w}$  are normalized. This method may give an error while processing normalized vectors. A natural method is adapted for non-normalized input. For this, winner neuron  $k$  with the corresponding weighting vector  $\mathbf{w}_k$  is

selected, which is closest to the input vector  $x$ , using the Euclidean distance measure:

$$k: \|w_k - x\| \leq \|w_o - x\| \quad \forall o \quad 4$$

It is easy to verify that if all the vectors are normalized, then (4) reduces to (1) and (2). The Euclidean distance norm is this more general case of formula (1) and (2). Instead of rotation of the weight vector in the direction of the input vector being performed according to formula (3), the weight can change using the shift in the direction of the input vector:

$$w_k(t+1) = w_k(t) + \gamma(x(t) - w_k(t)) \quad 5$$

Again only the weights of a neuron winner are changed.

The important step in this recursive procedure is initializing. If the input vectors form an input set of large dimension, then initialized at random weight vector  $w_o$  can never be elected as the winner and therefore, never to move and use.

Therefore, usually, weight vectors are initialized set of input vectors  $\{x\}$ , which are randomly, selected from the input set. An improved approach to avoiding this and other problems in the competitive learning is the so-called "leaky learning" [9]. The supplement changes weighting coefficients in accordance with (1.5) formula:

$$w_l(t+1) = w_l(t) + \gamma'(x(t) - w_l(t)) \quad \forall l \neq k \quad 6$$

The  $\gamma' \ll \gamma$  : coefficient «leaky learning.»

Another similar method is competitive training, sensitive to the frequency (frequency sensitive competitive learning) [11]. In this algorithm, for each neuron is memorized number of cases where he was elected as the winner. More the number of winning neurons, less susceptible will they be to competition.

### III. THE PROCESS OF CLASSIFICATION AND CLUSTERING OF MULTI-PARAMETRIC OBJECTS WITH THE SELF-ORGANIZING MAP OF KOHONEN

The artificial neural networks develop rules for a well-defined resolution, which allow classifying the multi-parametric objects (events, situations, processes). Unlike classical classification methods, the RNA is used to create a biological information model of the human and other beings brain [12, 13].

The model of the Kohonen network uses the competitive learning method [14]. This procedure distributes the objects of the learning multitude on clusters inherent to the regrouping of the input data [9]. During learning, the neurons are competing, and the network fixes the winner neurons for each group of similar input objects. The fixed neurons form the centers of the clusters. The metric used in this operation is the Euclidean distance between the synaptic weights vectors, and input objects vectors.

#### A. The structure of the Kohonen neural network

The Kohonen network is composed of two layers: the input layer and output layer – precisely SOM. The elements of the map are dispersed in a space - usually one-dimensional or two-dimensional (Figure .2). The input data is presented as a

matrix, the rows are the vectors of objects, and the columns are the components of these objects.

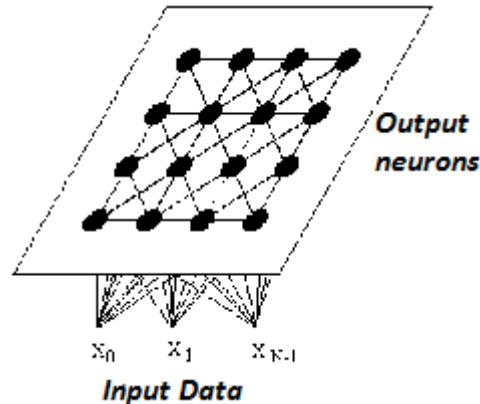


Fig. 2. Structure of the Kohonen map

The objects are presented to the network input one by one the idea of learning, is the application of successive rapprochement method, beginning with the random choice of the disposition of cluster centers, then the algorithm gradually improve them to perform the learning data clustering [7-8].

#### B. The Kohonen network learning algorithm

The learning procedure begins with the normalization of input data and synaptic weights to reduce the learning time [15]. This operation is based on the following algebraic formula:

$$x_i = x_i / \sqrt{\sum_{j=0}^{n-1} x_j^2} \quad 7$$

Where:  $x_i$  - input object component or of vector of synaptic weights;

$n$  - The number of variables in the vector  $x$ .

The main learning algorithm passes successively through a series of iterations. In each one, the learning object vectors are presented successively to network input, and the desired output is absent. At the end of this procedure, the topologically adjacent neurons respond to similar input vectors.

To fix the winner's neurons, using the metric of the Euclidean distance [4] see the formula below:

$$k: \|w_k - x\| \leq \|w_o - x\| \quad \forall o \quad 8$$

Subsequently the algorithm performs a correction of synaptic weights to gradually minimize the distance between the neuron winners and input objects. For this correction using the following formula [9]:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha_i(t)h(d, t) \cdot [y_i - w_{ij}(t)] \quad 9$$

Where:  $y_i$  - the value of the output neuron  $i$ ;

$w_{ij}(t)$  and  $w_{ij}(t+1)$  - the synaptic weights in iteration  $t$  and  $(t+1)$  iterations.

$\alpha_i(t)$ - learning rate, this coefficient can have a value between 0 and 1, and is calculated using the following equation:

$$\alpha_i = \alpha_0 e^{-\frac{t}{\mu}} \quad 10$$

Where: i - The iteration number;

t - The iteration rate.

$h(d, t)$  - neighborhood function, it is written according to the formula below:

$$h(d, t) = \begin{cases} 0, & d \geq \delta(t) \\ e^{-\frac{d}{2\delta(t)}}, & d < \delta(t) \end{cases} \quad 11$$

$$\delta(t) = \delta_0 e^{-\frac{t}{\mu}} \quad 12$$

Where: d- the distance between the winner neuron and an x neuron.

$\delta$  - Constant.

$$\mu = \frac{n}{\log_{10}(\delta_0)} \quad 13$$

n – Iteration rate.

The learning process will be continued up to the stabilization of the self-organizing map.

C. The classification and clustering process with Kohonen's self-organizing map: The experiments results

The test is performed to fix the impact of data normalization on the learning process, and the revelation of the restrictions which may limit the right application of the network in different domains.

The data used in the test are structured to form a matrix of a dimension of (8X3), whose lines present the input vector objects. Among these objects, there are two similar objects (Input 1 and 5), four objects having linear regularities between its components (Inputs 0, 4, 2 and 6) and two other normal objects (Inputs 3 and 7) see Tab.1.

The result presented in Figure 3 shows that the network designated a winner neuron for each pair of similar objects (1, 3, 5 and 7) and there are two different winner neurons for both normal objects which seem normal. However, for linear inputs objects (0, 2, 4 and 6), the network has fixed a single winner neuron for each pair that does not conform to our wishes.

TABLE I. INPUT VECTORS NORMALIZATION (MATRIX 8X3)

Inputs N°	Initial data	Data after normalization
	8 4 6	0,74 0,37 0,55
0	5 6 7	0,47 0,57 0,66
1	3 5 4	0,42 0,70 0,56
2	11 27 39	0,22 0,55 0,80
3	4 2 3	0,74 0,37 0,55
4	5 6 7	0,47 0,57 0,66
5	9 15 12	0,42 0,70 0,56
6	13 35 42	0,23 0,62 0,74
7		

	Input1	Input2	Input3	Input4	Input5	Input6	Input7	Input8
Inputs vectors	0.742781, 0.371391, 0.557086	0.476731, 0.572078, 0.667424	0.424264, 0.707107, 0.565685	0.225906, 0.554495, 0.800938	0.742781, 0.371391, 0.557086	0.476731, 0.572078, 0.667424	0.424264, 0.707107, 0.565685	0.231333, 0.622819, 0.747383
Winners vectors	0.742781, 0.371391, 0.557086	0.476731, 0.572077, 0.667424	0.424264, 0.707107, 0.565685	0.225708, 0.552759, 0.802368	0.742781, 0.371391, 0.557086	0.476731, 0.572078, 0.667424	0.424264, 0.707107, 0.565685	0.231349, 0.623026, 0.747221
Winners index	6	10	3	25	6	10	3	18
Map dimension	25							
Iterations rate	1000							
Iteration numbers	176							
Learning error	0,001							
Mean error	0,0006							
Neighborhood	2							
Radius rate	1,7715							

Fig. 3. Learning results of the self-organizing map Matrix 8X3

The theoretical analysis and experimental tests show that the possibilities of the card in the classification and Clustering task of multi-parameter objects are limited by the loss of certain information about objects (the initial length of vectors). This loss is caused by the need of the input vectors normalization. Thus, the practice shows that the speed and learning term depend on the size and complexity of the data multitude. Our perspective is to develop intelligent systems based on neural network models with competitive learning. These systems will be characterized with news capabilities to identify and classify correctly the linear or nearly linear objects.

IV. THE IMPROVED MODEL OF THE KOHONEN SELF-ORGANIZING MAP

The proposed new map model features an input data treatment and optimization block (IDOB). The proposed model structure is represented in the figure below.



Fig. 4. Structure of the new model of the Kohonen map

This block uses an algorithm combination of two very known and approved algorithms, the principal components analysis algorithm (PCA) [16], and the iterated power method IP [17]. This combination allows us to develop a method characterized by capacity derived from these two algorithms. The new skills eliminate the linear regularities between the objects components of the learning multitude, and reserve only the most informative components [18]. With these capabilities, the new implementation increases the performance and accuracy of classification and clustering.

A. The input data optimization algorithm

The proposed algorithm in this section combines two very known methods: the PCA that will calculate the correlation



matrix, and IP which will search the Eigenvectors that constitute the rows of the resulting matrix.

The principal component analysis is a way to identify patterns in the data, and expressing data so as to highlight their similarities and differences. The main advantage of the (PCA) is that once these patterns are found in the data, it can compress data by reducing the number of dimensions, without losing too much information. This technique is the basis of our method to achieve our objectives referred to in this work.

The initial data is organized as an array, whose rows are the individuals (objects) and columns present variables (Components) (see Figure5).

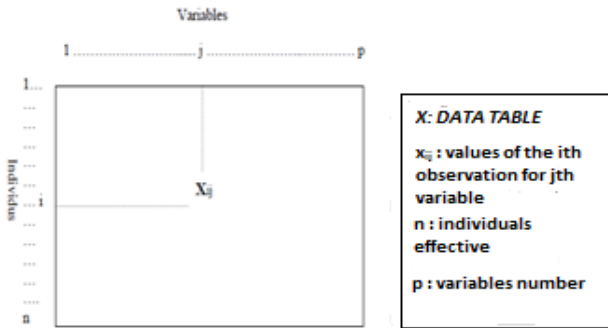


Fig. 5. The initial array data

In the first step, the algorithm calculates the vector of main point  $g$ . This point is the center of the points cloud in a space  $F$ . See the formula below.

$$g^t = (\bar{x}^1, \dots, \bar{x}^p) \tag{14}$$

$$\text{Où : } \bar{x}^j = \frac{1}{n} \sum_{i=1}^n x_i^j \tag{14.1}$$

At the base of the vector  $g$  is calculated the data centered matrix, which is written in terms of  $X$  of the following way:

$$Y = X - 1g^t \tag{15}$$

Where:  $g^t$  is the transposed of  $g$ .

The term centered signifies that the means of the variables  $\bar{y}^j$  are zero.

The centered data matrix  $Y$  is used in this step for calculating the variance-covariance matrix  $V$ , which is written according to the following way of  $Y$ :

$$V = \frac{1}{n} Y^t Y \tag{16}$$

Where:  $Y^t$  is the transposed of  $Y$ .

The  $V$  matrix is presented as follows:

$$V = \begin{pmatrix} S_{11} & \dots & S_{1p} \\ S_{21} & \dots & \dots \\ \dots & \dots & \dots \\ S_{p1} & \dots & S_{pp} \end{pmatrix} \tag{17}$$

Where:  $S_{kl}$  is the covariance of the variables  $k$  and  $l$ , and  $S_k$  is the variance of the variable  $k$ .

In the last step and to develop the correlation matrix  $R$ , we calculate the two diagonal matrices  $D_{1/S^2}$  et  $D_{1/S}$  as a function of  $V$  for the following way.

$$D_{1/S^2} = \frac{1}{\text{diag}(V)} \tag{18}$$

$$D_{1/S} = \frac{1}{\text{diag}(D_{1/S^2})} \tag{19}$$

The correlation matrix  $R$  comprises the coefficients of linear correlation between  $p$  variables taken two by two. It sums and shows the structure of linear dependencies between  $p$  variables. It is symmetrical and the diagonal is composed of 1.  $R$  is written according to the following way of  $V$ :

$$R = D_{1/S} V D_{1/S} \tag{4.7}$$

Where:  $D_{1/S}$  is the diagonal matrix containing  $\frac{1}{s_1}, \dots, \dots, \frac{1}{s_p}$  on its diagonal.

After calculating the matrix  $R$  using the iterated power method, the Eigenvectors are searched that constitute the rows of the resulting matrix  $M$  of input objects.

$$M = \begin{bmatrix} x_{11} & \dots & x_{1l} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nl} \end{bmatrix} \tag{20}$$

The figure below presents the proposed algorithm flowchart.

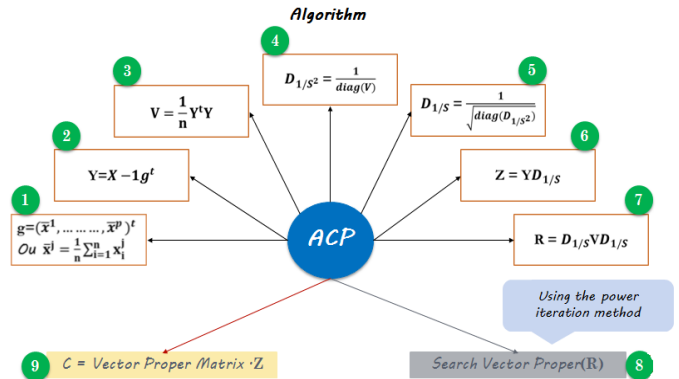


Fig. 6. the proposed algorithm flowchart

In the diagram, the first seven steps of the method ACP calculate the correlation matrix  $R$ . The next two steps of the iterated power method use this matrix to search the Eigenvectors and develop the reduced final matrix.

**B. The classification and clustering process with improved Kohonen's self-organizing map: The experiments results**

To test and verify the capabilities of the model proposed in comparison to the standard model, the experimental test in this section is based on the same data used to test the standard version, and with the same conditions. Thus, all the necessary modules have been added to our application, for which works in two modes: standard and modified. The figure below shows the result of the test performed.

	Input1	Input2	Input3	Input4	Input5	Input6	Input7	Input8
Inputs vectors	-0.65779	-1.02409	-1.53019	2.30907	-1.55335	-1.02409	0.307247	3.17319
Winners vectors	-0.65605	-1.02409	-1.53022	2.30908	-1.55337	-1.02409	0.307259	3.17318
Winners index	5	14	1	12	3	14	25	22
Map dimension	25							
Iterations rate	1000							
Iteration numbers	446							
Learning error	0,001							
Mean error	0,0007							
Neighborhood	2							
Radius rate	1,50004							

Fig. 7. the learning results of the improved model of map (Matrix 8X3)

The data in the figure above shows that the proposed model gave the desired results, and it has correctly classified the input objects. Based on these results, it is observed that the new model has designated seven winner neurons, unlike the standard model that has repartee the multitude of data only on five Clusters. Thus, we note that the IDOB has reduced the data size, and each object was described only by one component.

V. EVALUATION

To evaluate the innovations brought in this work and make conclusions, a graphical presentation of the input data figured in Table.2 is used and the results in the figures are given below.

TABLE II. INPUT VECTORS NORMALIZATION (MATRIX 4X2)

Initial Matrix	Matrix After Normalization
8 4	0,89 0,44
3 5	0,51 0,85
4 2	0,89 0,44
9 15	0,51 0,85

The length of the vectors used in this example (two components), has been adjusted to have the possibility of presenting input data into a two-dimensional space, which facilitates the understanding of the problems and improves results interpretation.

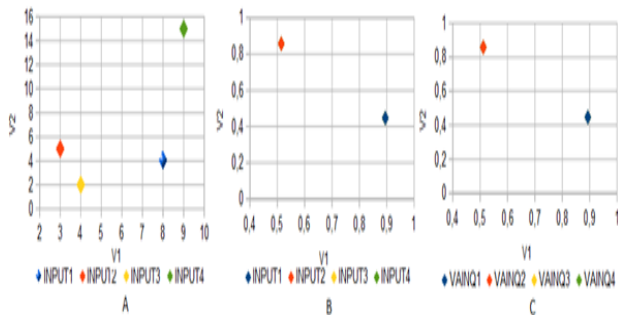


Fig. 8. Graphical presentation of data and the learning outcome of the SOM card (Classic Model)

The graphical presentation shows that the four input objects are correctly dispersed in the realization space (Figure 8.A), but after the normalization processing we can compress

the realization space but the vectors lose information about their initial lengths (Figure 8B). The loss of this information has prevented the standard model of the card correctly classify and clustering the data. Figure 8.C shows that after learning, the expected result was not obtained and the SOM has designated only two winner neurons for the four input objects.

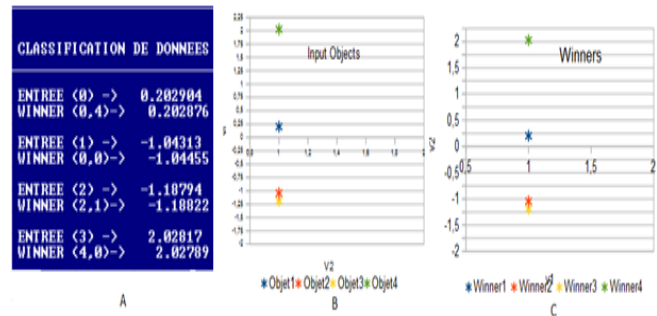


Fig. 9. Graphical presentation of input data and the SOM learning result (Improved Model)

In the second graphical presentation (Figure 9), we interpret the input data and results for the proposed model of the map. Part A of the figure shows that the IDOB has well filtered and reduced the initial data and the new dimension of the matrix formed by the block is (4x1). Thus, the three parts of the figure showing how the proposed model have eliminated the regularities between the components of objects and the learning algorithm has successfully brought the winners neurons to input objects.

The software developed for both map models (standard and modified); visualize the results on two-dimensional maps especially for the winners and their clusters (see both figures below). Map as a grid of rectangles corresponding to the neurons in the Kohonen layer, the index of the neuron in this layer is the same for the rectangle on the map.

To determine the winner neurons and Clusters on two maps the application uses the coloring technique. The rectangle that presents the winning neurons on the map will have a unique color; its neurons represent the centers of clusters in the second map. In the second map, the rectangles that present the same Cluster neurons will be colored with a unique color.

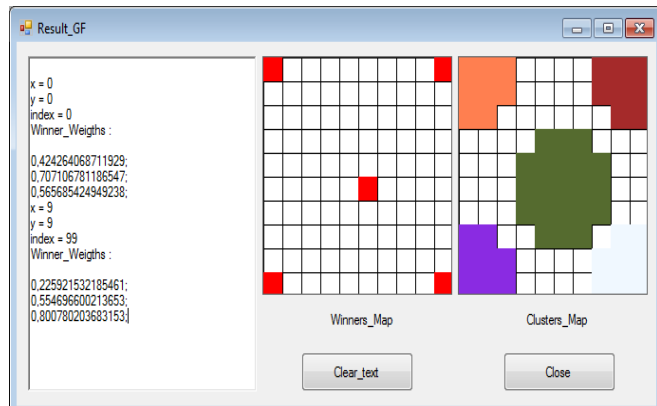


Fig. 10. the standard SOM, The results of the classification and clustering for the matrix (8 x 3)

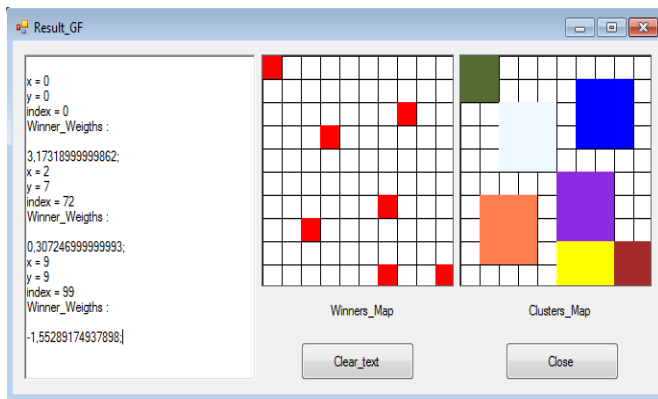


Fig. 11. The modified SOM, The results of the classification and clustering for the matrix (8 x 3)

According to the two figures, we can say that the proposed model exceeds the standard model concerning the accuracy of the classification. The two maps in Figure 10 show that the standard model has designated five winner neurons (five Clusters center) and five clusters, but maps in Figure 11 show that the proposed model has fixed seven centers of clusters according to the expected result.

## VI. CONCLUSIONS AND PERSPECTIVES

On the basis of evaluations and observations described above we arrive to the following conclusions and perspectives:

1) *The Kohonen network derived the animal brain functionality, and can be used to realize the intelligent systems.*

2) *The intelligent systems based on this paradigm are characterized by special abilities such as learning, adaptation and the possibility of visualization of Multiparameter objects with a reduced space.*

3) *The normalization procedure of input objects has advantages, but it prevents the card correctly classify linear or nearly linear objects.*

4) *The proposed method is approved theoretically. This method has eliminated the regularity problems between the components of linear objects and reduces the dimension of data.*

5) *The data processed by the blocks IDOB will be deformed, which complicates their analysis.*

6) *Despite its advantages, the proposed model of the card must be improved to solve the problem of the deformation of data, and increase the chance of its application in the different activity domains.*

## REFERENCES

- [1] M.M Glybovets, AV Olecko. Artificial Intelligence. - K. Publishing house "KM Academy", 2002. - 366 p.: fig. - Ref.: p. 339-351.
- [2] Tolstov Yu.N. Basics of multidimensional scaling. - M.: KDU, 2006. - 160 p. - ISBN 5-98227-100-4.
- [3] Joe. Kim, factorial, discriminate and cluster analysis, Edition - Ozon, 2012, 216p.
- [4] Gubler E.V. Computational methods of analysis and recognition of the pathological processes Leningrad: Medicina. 1978. 296c..

- [5] Hooper, Martyn (2013). "Richard Price, Bayes theorem, and God". Significance 10 (1):36–39. doi:10.1111/j.1740-9713.2013.00638.x.
- [6] I.P Gaydyshev Analysis and data processing: a Special reference book. - St. Petersburg: Peter, 2001. - 752 p.
- [7] K. Jane Anil, J. Mao, K.M Moinuddin, Introduction to Artificial Neural Networks // Open systems. - 1997. - №4. - p. 16-24.
- [8] A.A Ezhov, S.A Shumsky, Neurocomputing and its application in economics and business. - M.: MIFI, 1998. – 224.
- [9] Krose B., van der Smagt P. An introduction to neural networks. – The University of Amsterdam. – 1996. – 135 p.
- [10] Lippmann R.P. Review of neural networks for speech recognition // Neural Computation. – 1989. – №1. – P. 1-38.
- [11] Ahalt S.C., Krishnamurthy A.K., Chen P., Melton D. Competitive learning algorithms for vector quantization // Neural Networks. – 1990. – №3. – P. 277-290.
- [12] W.S McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys, 5, 115-133 (1943).
- [13] B.D Ripley, Pattern recognition and neural networks. – Cambridge: Cambridge University Press, 1996. – 403p.
- [14] T Kohonen, Self-organizing maps (2nd edition). – Springer Verlag, 1997. – 448p.
- [15] P. D. Wasserman. Neural Computing: Theory and Practice. ANZA Research, Inc. 1989. P 64.
- [16] Jérôme Pagès, Analyse factorielle multiple avec R, EDP sciences, Paris, 2013, 253 p.
- [17] Catherine Bolley. Analyse numerique. \_Ecole d'ingenieur. Nantes, France. 2012, p.97.
- [18] Michel Volle, Analyse des données, Economica, 4e édition, 1997.
- [19] R.Christian ;S.A.Yvan Mathématiques et technologie. Springer Science+ Business 2008. P 431.
- [20] M. Mc Cord Nelson, W.T. Illingworth A practical guide to neural nets, Addison-Wesley Publishing company 1991.
- [21] E. Davalo, P. Naïm, Des Réseaux de neurones, Edition Eyrolles 1993.
- [22] A.Deweze, L'accès en ligne aux bases documentaire, Collection MASSON 1983.
- [23] J. A. Farrel, A. N. Michel Associative memory via artificial neural networks IEEE control system magazine 1990.
- [24] J. A. Freeman, D. Skapma, Neural Networks, Algorithms, and programming technique, Addison-Wesley publishing company 1992
- [25] C.Guinchat, Y.Skouse, Guide pratique des techniques documentaires, Vol 1, 2, EDICEF 1989.
- [26] D. O. Hebb The organization of behavior J. Wiley & Sons NY, 1949 [HER 94] : J Herault, C Jutten, réseauxneurones et traitement de signal, Edition HERMES 1994.
- [27] C. L. JimenezGuaim, Opération d'accès par le contenu à une base de document textuelle, Thèse de doctorat INPG 1989.
- [28] T Kohonen, Self-Organized Formation of topologically correct features maps, Biological Cybernetics 1988.
- [29] TC. Leloup, Moteur d'indexation et de recherche, Edition Eyrolles, 1998.
- [30] Xia lin, D Soergel et G Marchionini, A self-organising semantic map for information retrieval, Int AMC/SIGIR Conference on R&D informationretrieval, pp 262-269, 1991
- [31] P. Martin Réseaux de neurones artificiels Application à la reconnaissance optique de partitions musicales Thèse de doctorat université de J. F. Grenoble I 1992.
- [32] P.Poincot, Création d'une carte bibliographique à l'aide des cartes autoorganisatrice de Kohonen, congrès INFOSID'97, Toulouse, 10-13 Juin 1997, pp 625-641.
- [33] G. N. Reddy, Artificial neural networks Theory, Computer simulation programs and Applications Dept. of Electrical Engineering Lamar University Beaumont, Texas, USA Nov. 1995.

# A New Approach for Time Series Forecasting: Bayesian Enhanced by Fractional Brownian Motion with Application to Rainfall Series

Cristian Rodriguez Rivero  
LIMAC – Department of Electronic Engineering  
Universidad Nacional de Córdoba  
Córdoba, Argentina

Julian Pucheta  
LIMAC – Department of Electronic Engineering  
Universidad Nacional de Córdoba  
Córdoba, Argentina

Daniel Patiño  
INAUT – Institute of Automatics  
Universidad Nacional de San Juan  
Cordoba, Argentina

Victor Sauchelli  
LIMAC – Department of Electronic Engineering  
Universidad Nacional de Córdoba  
Córdoba, Argentina

**Abstract**—A new predictor algorithm based on Bayesian enhanced approach (BEA) for long-term chaotic time series using artificial neural networks (ANN) is presented. The technique based on stochastic models uses Bayesian inference by means of Fractional Brownian Motion as model data and Beta model as prior information. However, the need of experimental data for specifying and estimating causal models has not changed. Indeed, Bayes method provides another way to incorporate prior knowledge in forecasting models; the simplest representations of prior knowledge in forecasting models are hard to beat in many forecasting situations, either because prior knowledge is insufficient to improve on models or because prior knowledge leads to the conclusion that the situation is stable.

This work contributes with long-term time series prediction, to give forecast horizons up to 18 steps ahead. Thus, the forecasted values and validation data are presented by solutions of benchmark chaotic series such as Mackey-Glass, Lorenz, Henon, Logistic, Rössler, Ikeda, Quadratic one-dimensional map series and monthly cumulative rainfall collected from Despeñaderos, Cordoba, Argentina. The computational results are evaluated against several non-linear ANN predictors proposed before on high roughness series that shows a better performance of Bayesian Enhanced approach in long-term forecasting.

**Keywords**—long-term prediction; neural networks; Bayesian inference; Fractional Brownian Motion; Hurst parameter

## I. INTRODUCTION

Forecasting is based on identifying and estimating through observation and in some instances theory, patterns and/or relationships and then extrapolating or interpolating them in order to predict [1]. Scientists are different than other forecasters in being well aware of uncertainty, by providing probabilistic forecasts, and constantly searching for enhancements using objective feedback [2]. Their overall success rate is improving and their assessment of uncertainty is well calibrated. Finally, weather forecasters have learned that predicting extreme weather events requires different models and skills than those of normal ones [3].

At the same time, they consider such events as an integral part of their job, even though it requires special effort, different models and extra skills to predict them.

The natural phenomena where humans cannot influence their future course, except to a limited extent, do influence with their actions and reactions, changing their future course, making forecasting more difficult but also more challenging [4]. The question regarding why simple forecasting models outperform sophisticated ones is still open. The future is never exactly like the past which means that the accuracy of extrapolative predictions cannot be assured. The crucial question is the extent of accuracy, or inaccuracy of such predictions. Most of the time series in rainfall forecast are influenced by random events and often behave not far from random walks, favoring simple methods that are capable of smoothing such randomness. In long-term forecasts, the accuracy of predictions drops while uncertainty increases. On way used by forecaster is reducing the forecasting errors of predictions by averaging more than one model [5]. The outcome is not only higher accuracy but also a reduction in the size of forecasting errors, with simple averaging being the best way of combining forecasts [6]. The reason is that averaging cancels out the errors of individuals and/or models and in doing so eliminates the noise from the pattern and improve accuracy.

Rossi, Allenby, and McCulloch (2006) argued that there are really no other approaches except the Bayesian approach which can provide a unified treatment of inference and decision as well as properly accounting for parameter and model uncertainty. The Bayesian approach allows researchers to cope with complex problems. However, the Bayesian inference provides answers conditional on the observed data rather than based on distribution of test statistics over imaginary samples not observed. Even though the Bayesian approach has decent benefits, it has some trivial costs including formulation of prior, requirement of a likelihood function, and computation of various integrals required in Bayesian paradigm [7]. The

advancement of computation make complicated integrals become possible. However, choosing an appropriate or objective prior has been an issue in the Bayesian approach [8]. Thus, investigators are facing a practical problem with little information in the real-world situations and should not neglect sources of information outside of the current data set [9].

In this article, the major advantage of the proposed BEA technique is that the complexity does not increase with an increasing number of inputs. The solutions can easily be generalized to the problem of uncertain (noisy) inputs, such as Bayesian inference [10] against other generalized approaches [11]. Here the filters in comparison are based on non-linear stochastic auto-regressive moving average (NAR) models such as Bayesian approach [12] and Neural-Network Modified [13] [14], implemented by ANN.

The paper is organized as follows: Section II presents the data series as an important case of study, such as well-known chaotic time series. Section III provides the Bayesian Enhanced approach as a method using fractional Brownian motion for obtaining optimal network model. In Section IV, the proposed prediction method is highlighted by showing the performance of the proposed algorithm detailing experimental setup, results and analysis, with Section V providing some discussions and concluding remarks.

## II. DATA TREATMENT

### A. Overview on fBm

The fractional Brownian motion, which provides a suitable generalization of the Brownian motion, is one of the simplest stochastic processes exhibiting long-range dependence [15] [16]. It has been used as a modeling tool. The following demonstrates the stochastic integral representation of fractional Brownian motion [17]. The process is as follows,

$$B(t) = \frac{1}{\Gamma\left(H + \frac{1}{2}\right)} \int_0^t \left( (t-s)^{H-\frac{1}{2}} - (-s)^{H-\frac{1}{2}} \right) dB(s) \quad (1)$$

$$= \frac{1}{\Gamma\left(H + \frac{1}{2}\right)} \left( \int_{-\infty}^0 \left( (t-s)^{H-\frac{1}{2}} - (-s)^{H-\frac{1}{2}} \right) dB(s) + \int_0^t \left( (t-s)^{H-\frac{1}{2}} - (-s)^{H-\frac{1}{2}} \right) dB(s) \right)$$

where  $B(t)$  is a standard Brownian motion and  $\Gamma$  refers to the gamma function, is a  $fBm$  with  $0 < H < 1$ . The constant  $1/\Gamma(H+1/2)$  in the following computation is dropped for the sake of simplicity. According to the definition, a fractional Brownian motion ( $B(H)(t)$ )  $t > 0$  of Hurst parameter  $H$  is a continuous and centered Gaussian process with covariance function,

$$B^H(s) = \frac{1}{2} \left( t^{2H} + s^{2H} - |t-s|^{2H} \right). \quad (2)$$

Therefore,  $B(t)$  is a  $fBm$  of Hurst index  $H$ . The  $fBm$  is divided into three different families corresponding to  $0 < H < 1/2$ ,  $H = 1/2$ ,  $1/2 < H < 1$ , respectively. The basic feature of  $fBm$  is that the span of independence between their increments can be infinite [18]. As the Hurst parameter  $H$  governs the fractal

dimension of the fractional Brownian motion, its regularity and the long-memory behavior of its increments, the estimation of  $H$  is an important but difficult task which has led to very vast literature [19].

In this work, the  $H$  index is measured by wavelet method [20] [21].

### B. Overview on Benchmark Time Series

The standard non-parametric approaches presented in this article are based on stochastic techniques that assume non-linear relationship among data that reproduce the benchmark chaotic time series and rainfall data only in statistical sense. Although there are many situations when accurate forecasting is impossible, there are many others where predictions can provide useful information to improve our decisions and gain from effective action. Weather forecasts, made several times a day, in hundreds of thousands of locations around the world, are an example, as it is proposed in this work. The rainfall dataset used is from Despeñaderos located at Cordoba, province of Argentina (-31.824703;-64.289692) and the collection date is from year 2000 to 2014 as shown in Fig.1.

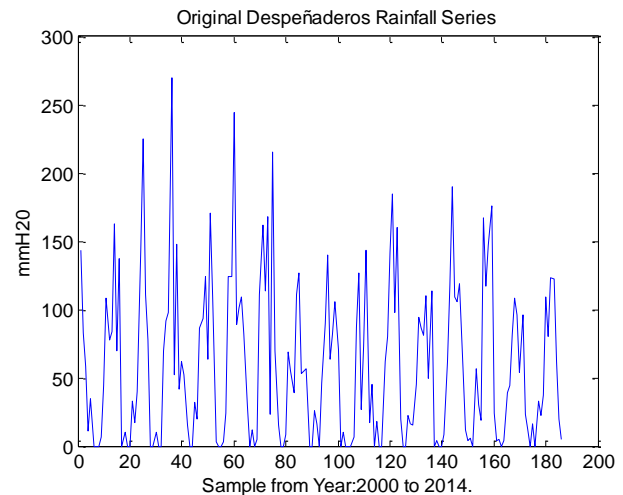


Fig. 1. Original Rainfall times series from Despeñaderos, Cordoba, Argentina

The rest of the benchmark time series are presented in the following subsections.

### C. The Mackay-Glass Chaotic Time Series

The dataset ensemble is by sampling the Mackay-Glass (MG) equations [22] defined by:

$$\dot{x}(t) = \frac{ax(t-\tau)}{1-x(t-\tau)^c} - bx(t) \quad (3)$$

with  $a, b, c, \tau$  setting parameters shown as follows in Table I.

TABLE I. PARAMETERS TO GENERATE MG TIME SERIES

Series No.	$\beta$	$\alpha$	$c$	$\tau$	$H$
MG17	0.2	0.1	10	17	0.980
MG30	0.2	0.1	10	30	1



D. The Logistic Chaotic Time Series

The logistic series (LOG) is defined by:

$$x(t+1) = ax(t)[1-x(t)]. \tag{4}$$

When a=4, the iterates of (4) perform a chaotic time series [23].

TABLE II. PARAMETERS TO GENERATE LOG TIME SERIES

Series No.	N	a	X <sub>0</sub>	H
LOG01	50	4	0.1	0.151
LOG03	50	4	0.3	0.100

E. The Henon Chaotic Time Series

The Henon chaotic time series can be constructed by following (5), however, it presents many aspects of dynamical behavior of more complicated chaotic systems [24].

$$x(t+1) = b+1-ax^2(t) \tag{5}$$

When generating data for our experiments, a and b are set as shown in Table III. These same parameters are used in both [25].

TABLE III. PARAMETERS TO GENERATE HEN TIME SERIES

Series No.	N	a	b	X <sub>0</sub>	Y <sub>0</sub>	H
HEN01	120	1.4	0.3	0	0	0.187
HEN03	120	1.3	0.22	0	0	0.833

F. The Lorenz Chaotic Time Series

Lorenz found three ordinary differential equations which closely approximate a model for thermal convection [26]. These equations have also become a popular benchmark for testing non-linear predictors. The Lorenz model is given by the equations (4), the data is derived from the Lorenz system, which is given by three time-delay differential systems

$$\begin{cases} \frac{dx}{dt} = a(y-x), \\ \frac{dy}{dt} = bx - y - xz, \\ \frac{dz}{dt} = xy - bz \end{cases} \tag{4}$$

A typical choice for the parameter values are as a = 10, b = 28, and c = 8/3. In this case, the system is chaotic. The data set is constructed by using four-order Runge–Kutta method with the initial value as is shown in Table IV for LOR01 and LOR03 series. The step size is chosen as 0.01, respectively. These sets of parameters are commonly used in generating the Lorenz system because exhibits deterministic chaos.

TABLE IV. PARAMETERS TO GENERATE ROS TIME SERIES

Series No.	n	X(0)	Y(0)	Z(0)	H
LOR01	120	12	9	2	0.158
LOR03	120	0.1	0.1	2	0.080

G. The Rössler Chaotic Time Series

In this example, the data is derived from the Rössler system [27], which is given by three time-delay differential systems. The data set is constructed by using four-order Runge–Kutta method with the initial value as shown in Table V, and the step size is chosen as 0.01, respectively.

$$\begin{cases} \frac{dx}{dt} = -y(t) - z(t), \\ \frac{dy}{dt} = x(t) + ay(t), \\ \frac{dz}{dt} = b + z(t)(x(t) - c) \end{cases} \tag{4}$$

TABLE V. PARAMETERS TO GENERATE LOR TIME SERIES

Series No.	n	a	b	c	X(0)	Y(0)	Z(0)	H
ROS01	120	0.2	0.2	5.7	12	9	2	0.203
ROS03	120	0.42	0.42	0.42	0.1	0.1	2	0.274

H. The Ikeda Chaotic Time Series

Before describing the reconstruction, I introduce the system which will be used to generate most of the time series described herein, namely the Ikeda map [28]. The Ikeda map is given as follows:

$$\begin{cases} \dot{x} = 1 - \mu[x \cos(t) - y \sin(t)], \\ \dot{y} = \mu[x \cos(t) + y \sin(t)] \end{cases} \tag{7}$$

where  $t = 1/(1+x^2+y^2)$ . This system displays chaotic behavior over a range of values for the parameter  $\mu$  including the values chosen here.

TABLE VI. PARAMETERS TO GENERATE IK TIME SERIES

Series No.	n	$\mu$	X(0)	Y(0)	H
IK01	120	0.9	0.5	0.5	0.029
IK03	120	0.8	0.9	0.6	0.060

I. The Quadratic Chaotic Time Series

The quadratic map is defined by the equation

$$x(t+1) = \mu x(t)(1-x(t)) \tag{7}$$

If this mapping is iterated by  $\mu=4$ , starting with a random number in the interval between 0 and 1, then different behavior is obtained dramatically depending upon the initial value of  $x$ . Initial values of  $x$  which are quite close together can have dramatically different iterates.

This unpredictability or sensitive dependence on initial conditions is a property familiar in displaying chaotic behavior [29] over a range of values for the parameter  $\mu$  including the values chosen here.

TABLE VII. PARAMETERS TO GENERATE QUA TIME SERIES

Series No.	n	c	X(0)	H
QUA01	120	1.95	0.5	0.31
QUA03	120	0.8	0.9	0.011

120 samples are for the selected time series, the first 102 values are used for training and the remaining 18 values are kept for validation and test data. The long-term behavior changes thoroughly by changing the initial conditions to obtain the stochastic dependence of the deterministic time series according to its roughness assessed by the H parameter. Then, extra 18 testing data are used to measure the prediction performance.



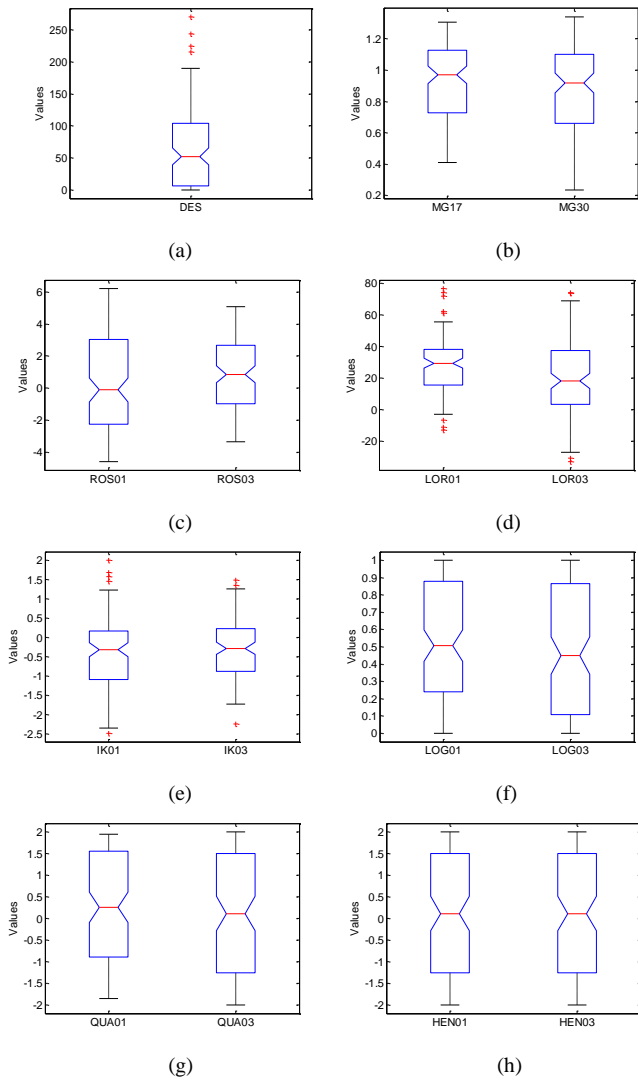


Fig. 2. Bloxplot diagram of the benchmark chaotic time series

### III. BAYESIAN ENHANCED APPROACH

This section presents a new method for issuing time series forecasting by focusing on three aspects: the formalization of one-step forecasting problems as supervised learning tasks, the discussion of modeling with Bayes inference techniques as an effective tool for dealing with temporal data and the key of the forecasting strategy when multiple-step-ahead is used for forecasting.

The increasing availability of large amounts of historical data and the need of performing accurate forecasting of future behavior in several scientific and applied domains demands the definition of robust and efficient techniques able to infer from observations the stochastic dependency between past and future. The forecasting domain has historically been influenced by linear statistical methods such as ARIMA models. More recently, machine learning models have drawn attention and have established themselves as serious contenders to classical statistical models in the forecasting community.

In this research, the Bayes assumption is used to update a prior distribution into a posterior distribution by incorporating

the information driven as likelihood function from fractional Brownian, provided by neural networks weights from observed data in order to generate point and interval forecasts by combining all the information and sources of uncertainty into a predictive distribution for the future values.

A weight vector  $w$  defines a mapping from an input vector  $x$  to a predicted output vector  $\hat{y}$  given by  $\hat{y} = f(x, w)$ . Assuming a fractional Brownian model, the conditional probability distribution for the output given the input vector  $l_x$  is as follows:

$$P(D / w, \mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{R}|^{\frac{1}{2}}} \exp\left(-\frac{(y(x_n; w) - t_n)^T \mathbf{R} (y(x_n; w) - t_n)}{2}\right), \quad (5)$$

$$P(t / H) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{R}|^{\frac{1}{2}}} \exp\left(-\frac{t^T \mathbf{R} t}{2}\right), \quad (6)$$

The application of the regression problem involving the correspond neural network function  $y(x, w)$  and the data set consisting of  $N$  pairs, input vector  $l_x$  and targets  $t_n$  ( $n=1, \dots, N$ ).

To complete the Bayesian enhanced approach for this work [30], prior information for the network is required. The beta distribution is chosen for this purpose. The Beta density function is a very versatile way to represent outcomes like proportions or probabilities. It is defined on the continuum between 0 and 1. There are two parameters  $\alpha$  and  $\beta$  which work together to determine if the distribution has a mode in the interior of the unit interval and whether it is symmetrical. This is a probability model which describes the knowledge gained after observing a set of data. It is proposed to use fractional Brownian, where  $H$  is the Hurst parameter,  $t_n = (t_1, t_2, \dots, t_n)$  follows a standard  $fBm$  and

$$R_{ij} = \frac{1}{2} (i^{2h} + j^{2h} - |i - j|^{2h}) \quad (7)$$

assuming that the expected scale of the weights is given by  $w$  set by hand. The Beta prior distribution for  $H$  is

$$P(h) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} h^{\alpha-1} (1-h)^{\beta-1}, \quad (8)$$

The full probability model is derived from the product,

$$P(w / h) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{R}|^{\frac{1}{2}}} \exp\left(-\frac{w^T \mathbf{R} w}{2}\right) \frac{\beta^\alpha}{\Gamma(\alpha)} h^{-\alpha-1} \exp\left(-\frac{\beta}{h}\right), \quad (9)$$

Then, the posterior distribution is as follows,

$$P(h / w) = \frac{\frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{R}|^{\frac{1}{2}}} \exp\left(-\frac{w^T \mathbf{R} w}{2}\right) P(h)}{\int_0^1 \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{R}|^{\frac{1}{2}}} \exp\left(-\frac{w^T \mathbf{R} w}{2}\right) P(h)}, \quad (10)$$

Metropolis-Hasting algorithm was used for computation with a starting value of 0.1. The number of iterations was set to be 10,000. Monte Carlo Error was used to examine the convergence. This was carried out considering that the network function  $f(x_n + I, w)$  is approximately linear with respect to  $w$  in the vicinity of this mode, in fact, the predictive distribution for  $y_{n+1}$  will be another multivariate Gaussian. This was carried out considering that the network function  $f(x_n + I, w)$  is

approximately linear with respect to  $w$  in the vicinity of this mode, in fact, the predictive distribution for  $y_{n+1}$  will be another multivariate Gaussian.

#### IV. PREDICTION RESULTS

The simulation results in different order approximations and time periods are presented in the following Table VIII. The performance of the comparison is measured by the Symmetric Mean Absolute Percent Error (SMAPE) and Root Mean Square Error (RMSE) proposed in the most of metric evaluations [31], defined by

$$SMAPE_s = \frac{1}{n} \sum_{i=1}^n \frac{|X_i - F_i|}{(|X_i| + |F_i|)/2} \cdot 100 \quad (11)$$

$$RMSE_s = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - F_i)^2} \quad (12)$$

where  $t$  is the observation time,  $n$  is the size of the test set,  $s$  is each time series,  $X_t$  and  $F_t$  are the actual and the forecasted time series values at time  $t$  respectively. The SMAPE and RMSE of each series  $s$  calculate the error in percent between the actual  $X_t$  and its corresponding forecast value  $F_t$ , across all observations  $t$  of the test set of size  $n$  for each time series  $s$ .

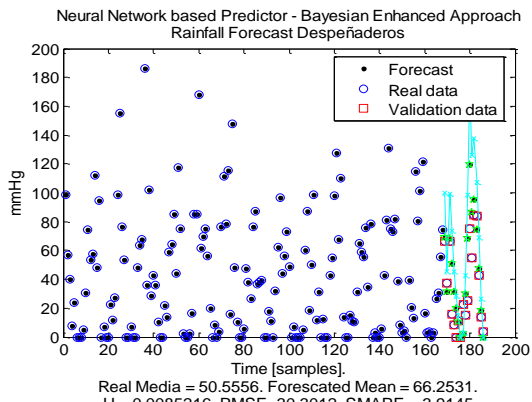


Fig. 3. Despeñaderos Rainfall time series Neural-Network Bayesian enhanced approach (BEA)

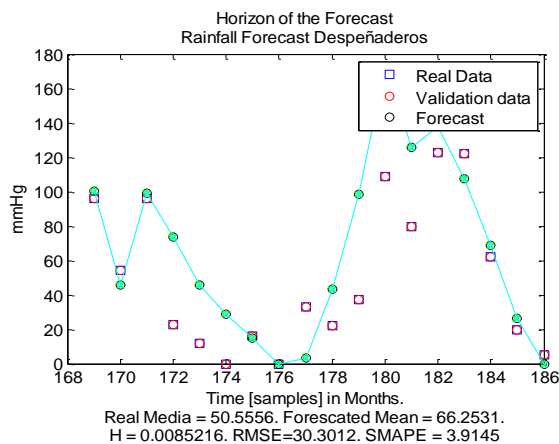


Fig. 4. Despeñaderos Rainfall time series Neural-Network Bayesian enhanced approach (BEA)

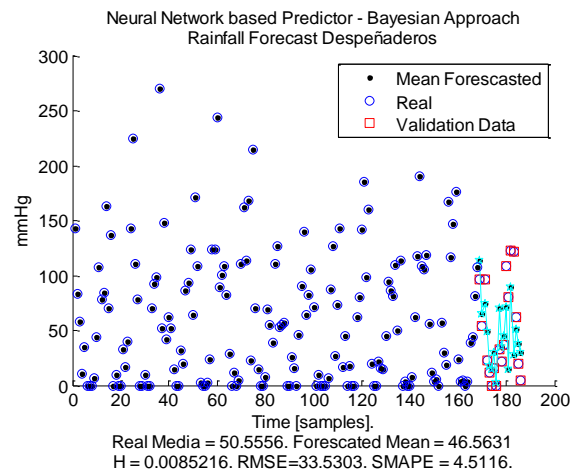


Fig. 5. Despeñaderos Rainfall time series Neural-Network Bayesian approach (BA)

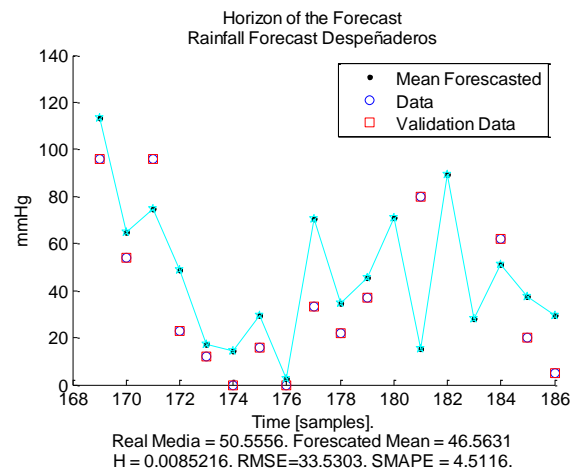


Fig. 6. Despeñaderos Rainfall time series Neural-Network Bayesian approach (BA)

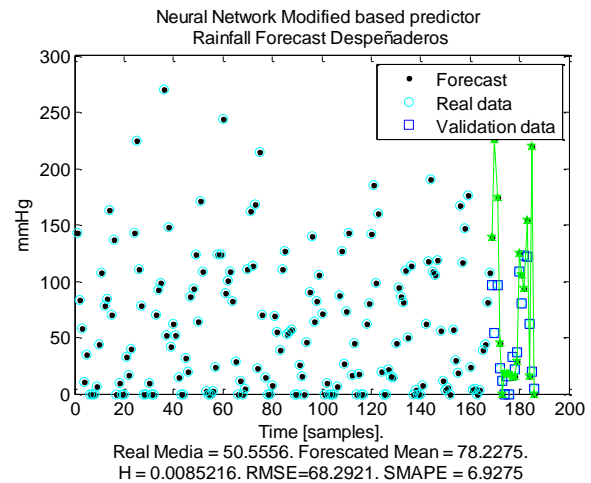


Fig. 7. Despeñaderos Rainfall time series Neural-Network Modified approach (NN-Mod)

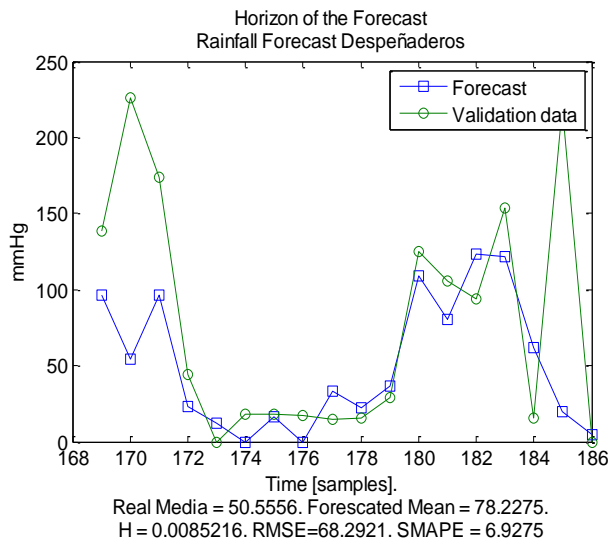


Fig. 8. Despeñaderos Rainfall time series Neural-Network Modified approach (NN-Mod)

The Monte Carlo method was used to forecast the next 18 values from benchmark chaotic and rainfall time series. Such outcomes are shown from Fig.3 up to Fig.8. Here, previous algorithms are used [10] [11] [13] [14] to compare the Bayesian enhanced approach.

Comparisons are performed between the Bayesian and NAR models by using long-term time series; in this case 15 years of monthly rainfall data (2000-2014) served as the historical data to forecast 2015 and benchmark of chaotic time series proposed in the literature. The results in Table VIII show that Bayesian enhanced approach (BEA) were a bit superior for the lengthy time series, with a SMAPE and RMSE about one half that of the Bayesian approach (BA) and neural network-modified predictor filter (NN-Mod).

The results show that the performances of the BEA with BA and NN-Mod are better than those in term of SMAPE and RMSE, due to the existence of outliers. With this lengthy series, BEA could adequately detect the underlying the relationship among those correlated variables. The simulation results of the BEA methods are compared with the BA and NN-Mod summarized in Table VIII for the benchmark time series. The similarity of the trend of the prediction performance between them is clear, BEMA is slightly better particularly on rainfall time series with reference to BEA and BA approaches.

TABLE VIII. RESULTS OF THE FORECASTING APPROCHES

Series	Method	Real Mean	Forecasted Mean	SMAPE	RMSE
Rainfall	BEA	50.55	66.25	3.91	30.30
	BA	50.55	46.56	4.51	33.53
	NN-Mod.	50.55	78.22	6.92	68.29
MG17	BEA	0.94	0.92	0.045	0.032
	BA	0.94	0.91	0.085	0.036
	NN-Mod.	0.94	0.89	0.32	0.047
MG30	BEA	0.87	0.85	0.027	0.09
	BA	0.87	0.91	0.11	0.017
	NN-Mod.	0.87	0.80	0.32	0.038
ROS01	BEA	0.11	0.10	0.009	0.05
	BA	0.11	0.14	0.017	0.10
	NN-Mod.	0.11	0.20	0.025	0.14
ROS03	BEA	0.0014	0.0026	0.00005	0.0003
	BA	0.0014	0.0030	0.00014	0.0014
	NN-Mod.	0.0014	0.0026	0.00031	0.0018
IK01	BEA	-0.44	-0.28	0.017	0.125
	BA	-0.44	0.008	0.105	0.29
	NN-Mod.	-0.44	0.012	0.136	0.99
IK03	BEA	-0.23	0.027	0.016	0.101
	BA	-0.23	0.053	0.021	0.121
	NN-Mod.	-0.23	0.017	0.159	0.93
LOR01	BEA	33.37	30.37	0.12	7.61
	BA	33.37	28.99	1.42	10.01
	NN-Mod.	33.37	17.79	5.06	32.41
LOR03	BEA	25.30	35.62	2.84	18.30
	BA	25.30	36.74	3.15	20.60
	NN-Mod.	25.30	23.30	4.95	30.08
LOG01	BEA	0.49	0.50	0.059	0.017
	BA	0.49	0.51	0.12	0.043
	NN-Mod.	0.49	0.46	0.35	0.063
LOG03	BEA	0.34	0.38	0.07	0.034
	BA	0.34	0.41	0.12	0.057
	NN-Mod.	0.34	0.47	0.40	0.072
QUA01	BEA	0.16	0.36	0.022	0.128
	BA	0.16	0.46	0.027	0.169
	NN-Mod.	0.16	0.53	0.22	1.32
QUA03	BEA	0.27	0.51	0.031	0.186
	BA	0.27	0.18	0.034	0.231
	NN-Mod.	0.27	0.04	0.27	1.51
HEN01	BEA	0.21	0.20	0.002	0.012
	BA	0.21	0.20	0.0025	0.014
	NN-Mod.	0.21	0.18	0.018	0.117
HEN03	BEA	0.13	0.13	0.0001	0.007
	BA	0.13	0.14	0.0019	0.012
	NN-Mod.	0.13	0.12	0.009	0.061

### V. DISCUSSION

This paper reports the results of a comparison of three different forecasting techniques for a class of high roughness long-term time series forecasting. The series were selected regarding the long or short term stochastic dependence of the time series assessed by the Hurst parameter  $H$  to give a forecast horizon of 18. The rainfall forecasts obtained between those algorithms are compared with NAR ANN predictor, namely BA and NN-Mod, for a case study on southwestern province of Cordoba. The study analyzed and compared the relative advantages and limitations of each time-series predictor filter technique, used for issuing rainfall and chaotic time series prediction. The discussion of how feed-forward networks can successfully approximate the quantitative changes in the dynamics of the time series data due to changes in the parameter values of the exogenous variables remains open for study, mainly.

Although the comparison was performed on ANN-based filters, the experimental results confirm that the enhanced Bayesian method can predict chaotic time series more effectively in terms of SMAPE and RMSE indices when is compared with other existing forecasting methods in the literature. However, the wish to preserve the stochastic dependencies constrains all the horizons to be forecasted with the same model structure. Since this constraint could reduce the flexibility of the forecasting approach, a variant of BEA approach is still open. Fig.10 and Fig.11 shows the evolution of the SMAPE and RMSE indices for BEA, BA and NN-Mod filter, which use the  $H$  parameter to adjust heuristically either structure of the net or parameters of the learning rule.

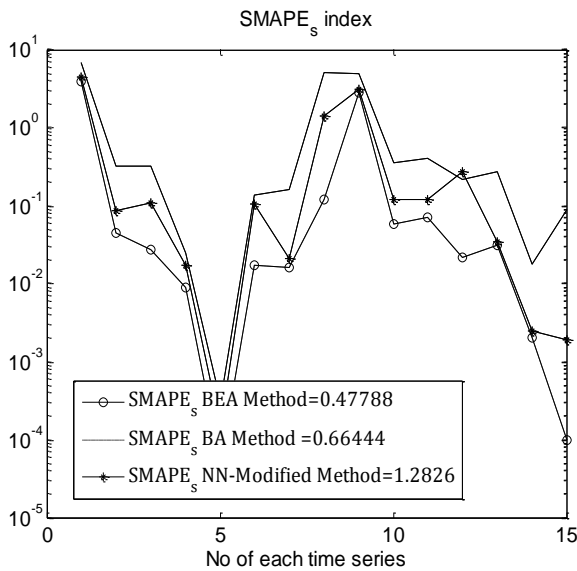


Fig. 9. The SMAPE index applied over the 15 time series

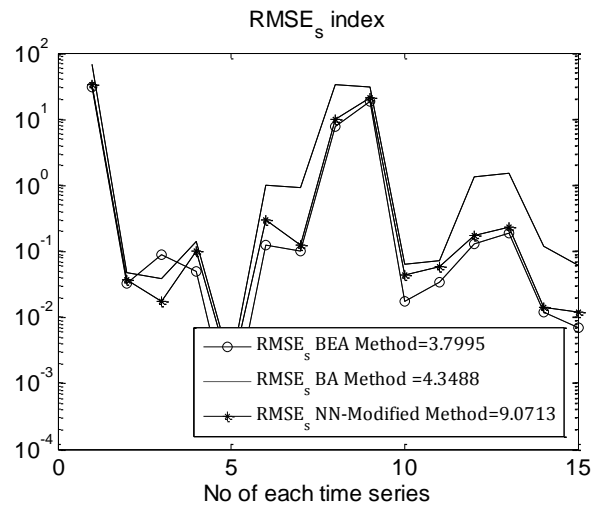


Fig. 10. The RMSE index applied over the 15 time series

### VI. CONCLUSIONS

A new approach for time series forecasting: Bayesian enhanced by fractional Brownian motion with application to rainfall series is presented. Building effective predictors from historical data demands computational and statistical methods for inferring dependencies between past and long-term future values of observed values as well as appropriate strategies to deal with longer horizons [32]. This work showed and discussed the BEA supervised learning technique to deal with long-term forecasting problems. In particular the  $fBm$  model assumption is stressed in Bayes inference by local learning approximators in dealing with important issues in forecasting, like nonlinearity and nonstationarity. The main results show a good performance in terms of SMAPE and RMSE indices of the predictor system based on Bayesian enhanced approach, particularly on rainfall time series from a geographical observation point, such as Despeñaderos, Cordoba, Argentina.

Future research should be concerned with the extension of these techniques to some recent directions in big data and the application to spatiotemporal tasks. These results encouraged us to continue working on new machine learning algorithms using novel forecasting methods.

### ACKNOWLEDGMENT

The authors wish to thank Agronomist Eng. Ernesto Carreño for providing monthly cumulative rainfall information at Despeñaderos, Cordoba, Argentina. The authors wish also to thank Universidad Nacional de Córdoba (UNC), Departments of Electrical and Electronic Engineering, FONCYT-PDFT PRH N°3 (UNC Program RRHH03), SECYT UNC, Institute of Automatics (INAUT) at Universidad Nacional de San Juan, National Agency for Scientific and Technological Promotion (ANPCyT) for their support of this work.

REFERENCES

- [1] Armstrong, J.S., Robert Fildes (2006). Making progress in forecasting, *International Journal of Forecasting*, Volume 22, No. 3.
- [2] Green, K. C. and Armstrong, J. S. (2015) *Journal of Business Research*, Volume 68, Issue 8, Pages 1657-1818 (August 2015) "Special Issue on Simple Versus Complex Forecasting".
- [3] Buchanan, M. (2013). *Forecast: What Physics, Meteorology, and Natural Sciences Can Teach us about Economics*, New York: Bloomsbury.
- [4] Fildes, R., and Makridakis, S., (1995). "The Impact of Empirical Accuracy Studies on Time Series Analysis and Forecasting", *International Statistical Review*, Vol. 63, no. 3, p. 289-308.
- [5] Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography, *International Journal of Forecasting*, No. 5, p. 559-583.
- [6] Makridakis, S., and Winkler, R. (1983). "Average of Forecasts: Some Empirical Results", *Management Science*, Vol. 29, No. 9, p. 987-996.
- [7] Rossi, P. E., Allenby G. M., & McCulloch, R. (2006). *Bayesian statistics and marketing*. Hoboken, NJ: John Wiley & Sons.
- [8] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). New York, NY: Chapman & Hall/CRC.
- [9] Chatfield, C. (2004). *The analysis of time series: An introduction* (6th ed.). London, England: Chapman and Hall.
- [10] Cristian M. Rodríguez Rivero, Julián A. Pucheta. FORECASTING RAINFALL TIME SERIES WITH STOCHASTIC OUTPUT APPROXIMATED BY NEURAL NETWORKS BAYESIAN APPROACH. *International Journal of Advanced Computer Science and Applications (IJACSA)*. West Yorkshire: The Science and Information (SAI) Organization. 2014 vol.5 n°6. p 145 - 150. ISSN 2156-5570.
- [11] Julián Pucheta, C. Rodríguez Rivero, Martín Herrera, Carlos Salas, Víctor Sauchelli, H. Daniel Patiño, Chapter: NON-PARAMETRIC METHODS FOR FORECASTING TIME SERIES FROM CUMULATIVE MONTHLY RAINFALL. *Book Rainfall: Behavior, Forecasting and Distribution*. Editors: Olga E. Martín and Tricia M. Roberts, Nova Science Publishers, Inc. ISBN: 978-1-62081-551-9 (2012).
- [12] Cristian M. Rodríguez Rivero, Julián A. Pucheta, Martín R. Herrera, Víctor Sauchelli, Sergio Laboret. Time Series Forecasting Using Bayesian Method: Application to Cumulative Rainfall, (Pronóstico de Series Temporales usando inferencia Bayesiana: aplicación a series de lluvia de agua acumulada). ISSN 1548-0992. Pp. 359 364. *IEEE LATIN AMERICA TRANSACTIONS*, VOL. 11, NO. 1, FEB. 2013.
- [13] Pucheta, J., Patiño, H.D., Kuchen, B. (2007). Neural Networks-Based Time Series Prediction Using Long and Short Term Dependence in the Learning Process. In *proc. of the 2007 International Symposium on Forecasting*, New York, USA.
- [14] Pucheta, J., Patiño, H. D., Kuchen, B. "A Statistically Dependent Approach for the Monthly Rainfall Forecast from One Point Observations". In *Proc. of the Second IFIP Conference on Computer and Computing Technologies in Agriculture (CCTA2008)* October 18-20, 2008. Beijing, China. (2008).
- [15] Bishwal, J. P. N. (2003). Maximum likelihood estimation in partially observed stochastic differential system driven by a fractional Brownian motion. *Stochastic Analysis and Applications*, 21(5), 995-1007.
- [16] Breton, A. L. (1998). Filtering and parameter estimation in a simple linear system driven by a fractional Brownian motion. *Statistics and Probability Letters*, 38, 263-274.
- [17] Biagini, F., Hu, Y., Øksendal, B., & Zhang, T. (2010). *Stochastic calculus for fractional Brownian motion and applications*. Berlin: Springer.
- [18] Mandelbrot, B. B., & Van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10, 422-437.
- [19] Achard, S., & Coeurjolly, J. F. (2010). Discrete variations of the fractional Brownian motion in the presence of outliers and an additive noise. *Statistical Surveys*, 4, 117-147.
- [20] Abry, P.; P. Flandrin, M.S. Taqqu, D. Veitch., Self-similarity and long-range dependence through the wavelet lens. *Theory and applications of long-range dependence*, Birkhäuser, pp. 527-556. 2003.
- [21] Bardet, J.-M.; G. Lang, G. Oppenheim, A. Philippe, S. Stoev, M.S. Taqqu. Semi-parametric estimation of the long-range dependence parameter: a survey. *Theory and applications of long-range dependence*, Birkhäuser, pp. 557-577. 2003.
- [22] M. Mackey and L. Glass, \Oscillation and chaos in physiological control systems," *Science*, vol. 197, p. 287, 1977.
- [23] A. Lapedes and F. Farber. *Nonlinear signal processing using neural networks: Prediction and System modelling*. LA-UR 87-2662. Los Alamos National Laboratory, 1987.
- [24] H. Abarbanel, *Analysis of Observed Chaotic Data*. 1996.
- [25] P. Cortez, M. Rocha, and J. Neves, \Evolving time series forecasting neural network models," 2001.
- [26] A. Gholipour, B. Araabi, and C. Lucas, \Predicting Chaotic Time Series Using Neural and Neurofuzzy Models: A Comparative Study," *Neural Processing Letters*, vol. 24, pp. 217-239, 2006.
- [27] Characterization of Strange Attractor. Grassberger, Peter and Procaccia, Itmar. s.l.: *Physical Review Letters*, 1983, Vols. Volume 50, Number 5.
- [28] K. Ikeda. *Opt. Commun.* 30, 257 (1979) .
- [29] Sprott, J. C. "Automatic Generation of Strange Attractors." *Comput. & Graphics* 17, 325-332, 1993. Reprinted in *Chaos and Fractals, A Computer Graphical Journey: Ten Year Compilation of Advanced Research* (Ed. C. A. Pickover). Amsterdam, Netherlands: Elsevier, pp. 53-60, 1998.
- [30] D. J. C. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural Comput.*, vol. 4, no. 3, pp. 448-472, May 1992.
- [31] Armstrong, J. S. (2001). *Forecasting Principles*. 2001.
- [32] Mendoza, M. & de Alba, E. (2006). Forecasting an accumulated series based on partial accumulation II: A new Bayesian method for short series with seasonal patterns, *International Journal of Forecasting*, Issue 4, 781-798.

# Performance Evaluation of Content-Based Image Retrieval on Feature Optimization and Selection Using Swarm Intelligence

Kirti Jain<sup>1</sup>,

<sup>1</sup>(Computer Science & Engg Dept., LNCT GROUP  
BHOPAL, INDIA)

Dr.Sarita Singh Bhadauria<sup>2</sup>

<sup>2</sup>(Electronics Dept., M.I.T.S.  
GWALIOR, INDIA)

**Abstract**—The diversity and applicability of swarm intelligence is increasing everyday in the fields of science and engineering. Swarm intelligence gives the features of the dynamic features optimization concept. We have used swarm intelligence for the process of feature optimization and feature selection for content-based image retrieval. The performance of content-based image retrieval faced the problem of precision and recall. The value of precision and recall depends on the retrieval capacity of the image. The basic raw image content has visual features such as color, texture, shape and size. The partial feature extraction technique is based on geometric invariant function. Three swarm intelligence algorithms were used for the optimization of features: ant colony optimization, particle swarm optimization (PSO), and glowworm optimization algorithm. Coral image dataset and MatLab software were used for evaluating performance.

**Keywords**—CBIR; Swarm intelligence; feature extraction; SIFT transform; GSO (glowworm swarm optimization)

## I. INTRODUCTION

Content-based image retrieval has started playing a major role in multimedia data storage and retrieval. A large amount of multimedia data is generated in today's times, such as image and video carton. For the purpose of online retrieval of these data, content-based image retrieval is used. The major issue in this process is precision and recall of retrieval image according to the query image [1,2]. For improvement of precision and recall, various authors used feature optimization and feature selection technique. The feature selection and feature optimization technique improved the performance of content-based image retrieval. For the extraction of feature, various pixel-based techniques and transform-based techniques were used. For transform-based technique, Wavelet transform function, Gabor transform function, SIFT transform function, and many other transform function-based techniques were used. In this paper, partial feature extraction technique was used, which overcomes the limitation of the other feature extraction techniques and improves the quality of partial feature [10]. The partial feature extraction technique is based on a geometrical invariant function such as sine, cosine, and tangent function and creates shape and size feature [5]. The feature extraction process gives a large amount of feature in terms of lower content feature and higher content of feature. Most of the authors do not consider the lower content of feature. Feature optimization technique was used for the lower content of feature. The feature optimization technique

optimized the optimal feature required for the process of feature extraction. In this paper, a swarm intelligence based feature optimization technique is used. The swarm intelligence family consists of various algorithms such as particle swarm optimization, ant colony optimization, and glowworm algorithm. These three algorithms are used for the process of feature optimization. The particle swarm optimization technique is a well-known dynamic population based optimization technique. The concept of particle swarm intelligence is based on bird flocks. The bird flocks maintain the continuous velocity and speed for flying and define two parameters namely Gbest and Pbest. The ant colony optimization technique works on the nature of biological ants. The biological ants find the path nearest to food. The glowworm optimization algorithm is especially useful for a simultaneous search of multiple optima, usually having different objective function values [9]. To achieve this goal, a swarm must have the ability to split into disjointed groups. Otherwise, only one (local or global) optimum will be found. In GSO, agents exchange information locally. Moreover, their movements are non deterministic. Section II of this paper describes partial feature extraction technique. Section III discusses feature optimization technique. Section IV discusses the experimental result analysis and Section V is the conclusion.

## II. PARTIAL SHAPE FEATURE EXTRACTION

This section discusses the feature extraction technique. The feature extraction technique is a very important phase of content-based image retrieval. The partial feature extraction technique has been used for feature extraction. The partial feature extraction technique was derived by Kirti Jain in 2013 [10]. The conventional shape-based feature extraction, such as chain code, edge detection, and Hough transform function are used for outer boundary feature detection. When the shape of the image used is of triangular and trapezoidal pattern the extraction of feature process such as chain code and edge detection suffer. Hence, some authors used ringlet transform function [3] for resolution of point function. However, the computation of point function is very complex, and so is the feature extraction process. The sine function, cosine function and tangent function are used for partial feature extraction based on boundary value of an image. The given image is divided into three sections namely hypotenuse, opposite, and adjacent. These three parameters are obtained before applying the edge detection technique for getting the X and Y



parameter in the plane. For better continuity of edge detection, the canny edge detection is used. The derivation process is now explained as formulae.

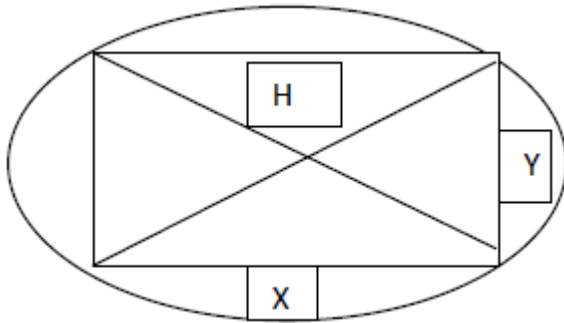


Fig. 1. Circular image divided into three sections: X, Y and H

The following steps explain the process of feature extraction using the triangular formula.

- 1) Apply canny edge detection method for finding the boundary value of the image
  - 2) Find the centred point of boundary value of shape
  - 3) Find the Xc and Yc as
 
$$Xc = \sum_{i=1}^n xi/n \dots \dots \dots (1)$$

$$Yc = \sum_{i=1}^n yi/n \dots \dots \dots (2)$$
  - 4) After getting a value of (Xc and Yc)
 
$$H = \sqrt{a^2 + b^2} \text{ Where } a = \sum_{i=1}^n \frac{xi}{n}, b = \sum_{i=1}^n \frac{yi}{n}$$
  - 5) After getting the value of H apply sine, cosine, and tangent function for shape of the boundary
  - 6)  $\text{Sin} = Xc/H$  and  $\text{cosine} = Yc/H$  and  $\text{tangent} = Yc/Xc$
  - 7) After getting sin, cosine, and tangent, find three consecutive matrix of shape
  - 8) All three shape parameters match the boundary value of the feature.
- This is the basic principle component of partial shape feature extraction process in image retrieval.

### III. FEATURE OPTIMIZATION PROCESS

For the optimization of feature, three swarm intelligence algorithms namely particle swarm optimization, ant colony optimization, and glowworm optimization algorithm are used. All these optimization algorithms are used in three different sections. The first section discussed the particle swarm optimization technique in terms of feature optimization; the second section discusses ant colony optimization technique in terms of feature optimization; and the third section discusses the glowworm optimization algorithm for feature optimization.

#### A. First Section

The particle swarm optimization feature optimizer selects the input of image feature (partial feature) in terms of  $D^R$  where D is the value of domain database and R shows the feature component of R image. The R image feature content  $\{r1, r2, r3, r4, \dots, rn\}$  describes the artificial particle as population. The unique feature relation of input image sets the velocity of the particle. If the feature attribute value is

changed, then the next iteration moves the update of velocity. These terms describe the particle's feature value; Rid and its near value of a particle; and Rgd which is the velocity value of optimization feature space. The random values for feature are fet1 and fet2 which are used for the local and global value selection of particle, that is, to make the optimal solution. The values of c1 and c2 manage the value of velocity of Rid and Rgd in deciding the particle's next movement velocity. Each iteration changes the velocity of swarm and creates a new feature subset for selection of feature. The derivation of equation in (3) and (4) [9]

$$v_{id} = w \times v_{id} + c1 \times fet1 \times (R_{id} - x_{id}) + c2 \times rand2 \times (R_{gd} - x_{id}) \dots \dots \dots (3)$$

$$X_{id} = x_{id} + v_{id} \dots \dots \dots (4)$$

Where w denotes the value of feature matrix; Rid is the position of particle value, Rgd is the position of global value best fitness value, c1 and c2 are constants and are known as acceleration coefficients; d denotes the dimension of the problem space; and fet1, fet2 are random values in the range of (0, 1).

#### B. Second Section

The extracted partial feature passes through ant colony optimization. The ant colony optimization process finds the continuity of shape feature. The process of ant colony optimization technique is basically described in terms of artificial ants. This process finds the dissimilar and redundant group of partial features [6]. The process of feature optimization is described here. The process of feature optimization of partial feature data passes through the feature space of ant colony optimization. The mapping of partial data feature attributes according to their artificial ants requires some standard derivation and parameter. On the basis of the parameters, the feature similarity of two different shapes is estimated. The most similar features pass through the process of retrieval and increase the capacity of precision and recall.

When F is a feature set and N is the total artificial ants and possibility of ant selection is  $s1, s2, \dots, sn$ , the selection possibility of two ants in giving solution is as follows.

$$SP(i, j) = \frac{1}{si - sj} \dots \dots \dots (5)$$

Where si and sj are the dissimilar probability of two different ants. Now the value of apptence of ants is estimated as

$$ACP(i+j) = \frac{\alpha i + \beta j}{N} \dots \dots \dots (6)$$

Where  $\alpha i$  and  $\beta j$  are ants whose selection possibility is maximum in terms of other ants; the ratio of selection of ants is defined as  $\frac{100}{N}$

On the basis of selection possibility, the value of artificial phenomenon value is estimated as follows.

$$\Delta \tau i = \frac{A \cdot si}{ACP(i+j)} \dots \dots \dots (7)$$

Where A is a constant phenomenon value

Now each iteration of pheromone value is increment and decrement according to their selection probability. The derivation of universal appetence probability is

$$p_{ij}^k = \begin{cases} \tau_{ij(t)} & \\ 0 & \text{otherwise} \end{cases} \alpha \cdot [k_{ij}] \beta i f j \epsilon j \dots \dots \dots (8)$$

Where  $k_{ij}$  gives the information of heuristic search space and measures the selection possibility of artificial ants

Finally, we get the optimal partial feature of image database for the purpose of optimization.

C. Third Section

The extracted partial feature passes through the glowworm algorithm. The partial feature is mapped into the glowworm search space. Each glowworm encodes the object function value  $J(x_i(t))$  at its current location  $x_i(t)$  into  $\alpha$  luciferin value  $l_i$  and broadcasts the same within its neighborhood. The set of neighbor ( $N_i(t)$ ) of glowworm  $i$  consists of those glowworms that have a relatively higher luciferin value located within a dynamic decision domain and updating by formula 4.1 at each iteration[11].

Local decision range update is given by equation “9”

$$r_d^i(t + 1) = \min \{rs, \max\{0, r_d^i(t) + \beta(nt - |N_i(t)|)\} \dots \dots \dots (9)$$

And  $r_d^i(t + 1)$  is the glowworm is local decision range at the  $t+1$  iteration,  $rs$  is the sensor range, and  $nt$  is the neighborhood range. The number of glow in local decision range is given by equation “10”

$$N_i(t) = \{j: \|x_j(t) - x_i(t)\| < r_d^i; l_j(t) > l_i(t)\} \dots \dots \dots (10)$$

$x_i(t)$  is the glowworm  $i$  position at the  $t$  iteration;  $l_i(t)$  is the glowworm luciferin at the  $t$  iteration. The set of neighbors of glowworm consists of those glowworms that have relatively higher luciferin value and are located within the dynamic decision domain whose range  $r_d^i$  is bounded above by a circular sensor range.

Each glowworm is given in equation “11”

$$p_{ij(t)} = \frac{l_j(t) - l_i(t)}{\sum_{k \in N_i(t)} l_k(t) - l_i(t)} \dots \dots \dots (11)$$

Movement update is given in equation “12”

$$x_{i(t+1)} = x_i(t) + s \left( \frac{s_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|} \right) \dots \dots \dots (12)$$

Luciferin update is given in equation “13”

$$l_i(t) = (1 - \rho) l_i(t-1) + \gamma j(x_i(t)) \dots \dots \dots (13)$$

And  $l_i(t)$  is a luciferin value of glowworm  $i$  at the  $t$  iteration,  $P$  belong  $(0,1)$  lead to the reflection of the

cumulative kindness of the path followed by the glowworm in their current luciferin values the parameter  $\gamma$  only scale the function values,  $J(x_i(t))$  is the value of test function. Finally, the optimal feature is obtained. The optimal feature of image database passes through the image retrieval process.

IV. RESULT ANALYSIS OF IMAGE RETRIEVAL

This section discusses the performance evaluation of three swarm intelligence algorithms for content-based image retrieval. For the evaluation of performance, coral image dataset 10000 was used. This coral image dataset was divided into different dataset models in terms of 1000, 2000, 5000, and 10000 and MatLab software was used for implementing the swarm-based algorithm. MatLab is a well-known computational algorithmic software. For the evaluation of performance, precision and recall were used as the two parameters. For the calculation of precision and recall, the following parameters were used.

TP- True positive

TN True Negative

FP False Positive

FN False Negative

$$Precision = \frac{TP}{(TP + FP)} \dots \dots \dots (a)$$

$$Recall = \frac{TP}{(TP + FN)} \dots \dots \dots (b)$$

Formula (a) and (b) measure the performance of precision and recall during the process of experimental evaluation.

TABLE I. COMPARATIVE PRECISION AND RECALL OF IMAGE RETRIEVAL BASED CLASSIFICATION FOR A TOTAL OF 1000 IMAGES

Data set	Method	Precision (%)	Recall (%)
Car images	GSO	88.66	82.22
	ANT	89.33	85
	PSO	91.26	87
Bird images	GSO	91.26	93.05
	ANT	92.48	94.06
	PSO	94.58	95.68
House images	GSO	86.33	80.81
	ANT	88.79	81.34
	PSO	90.76	83.67
Cartoon images	GSO	85.33	78.83
	ANT	91.33	81.33
	PSO	93.67	85.89
Multiple bird images	GSO	88.66	79.66
	ANT	92.49	80.68
	PSO	94.38	87.94

TABLE II. COMPARATIVE PRECISION AND RECALL OF IMAGE RETRIEVAL BASED CLASSIFICATION FOR A TOTAL OF 2000IMAGES

Data set	Method	Precision (%)	Recall (%)
Car images	GSO	85.89	82.22
	ANT	87.33	84.67
	PSO	93.56	88.79
Bird images	GSO	89.78	91.35
	ANT	93.88	92.16
	PSO	94.64	94.68
House images	GSO	87.46	81.76
	ANT	90.49	83.44
	PSO	92.86	85.72
Cartoon images	GSO	88.43	80.23
	ANT	92.63	83.53
	PSO	95.22	89.79
Multiple bird images	GSO	89.36	80.46
	ANT	91.46	82.36
	PSO	95.79	84.48

TABLE III. COMPARATIVE PRECISION AND RECALL OF IMAGE RETRIEVAL BASED CLASSIFICATION FOR A TOTAL OF 5000 IMAGES

Data set	Method	Precision (%)	Recall (%)
Car images	GSO	86.97	84.78
	ANT	89.43	85.7
	PSO	94.52	91.46
Bird images	GSO	90.48	92.5
	ANT	92.86	93.6
	PSO	95.66	94.66
House images	GSO	89.47	83.79
	ANT	91.89	85.54
	PSO	93.46	87.78
Cartoon images	GSO	89.43	82.23
	ANT	91.69	85.7
	PSO	93.62	89.79
Multiple bird images	GSO	91.65	80.86
	ANT	93.4	82.44
	PSO	95.7	84.5

TABLE IV. COMPARATIVE PRECISION AND RECALL OF IMAGE RETRIEVAL BASED CLASSIFICATION FOR A TOTAL OF 10000 IMAGES

Data set	Method	Precision (%)	Recall (%)
Car images	GSO	89.47	88.6
	ANT	91.63	89.4
	PSO	94.02	93.66
Bird images	GSO	91.48	93.5
	ANT	92.76	94.62
	PSO	95.48	95.7
House images	GSO	91.7	83.86
	ANT	93.8	85.5
	PSO	95.56	87.7
Cartoon images	GSO	91.3	82.28
	ANT	93.6	85.76
	PSO	95.68	89.8
Multiple bird images	GSO	93.69	80.59
	ANT	95.46	82.23
	PSO	96.8	84.57

Comparative result of 1000 images using GSO, ANT and PSO method

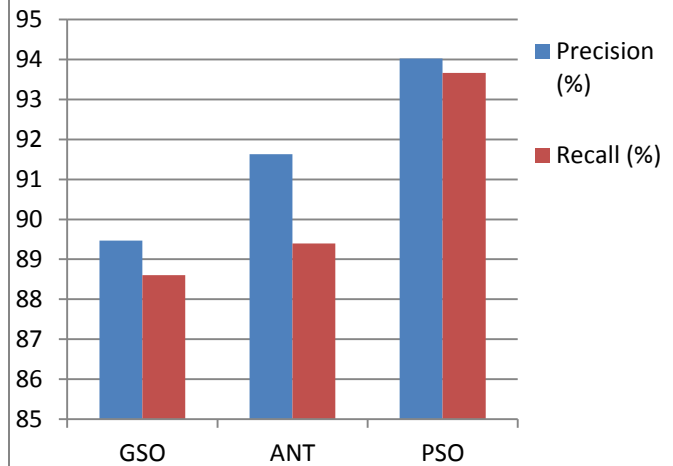


Fig. 2. Shows performance of a data set for 1000 image counts of data and rate of precision and recall on the base of GSO, ANT and PSO method. The above graph shows the comparative result analysis of image data sets of 1000 images with three different methods and calculates the parameter value of precision and recall. The variation of result in precision and recall on the basis of applying three methods found that the result of PSO is better than GSO and ANT

Comparative result of 2000 images using GSO, ANT and PSO method

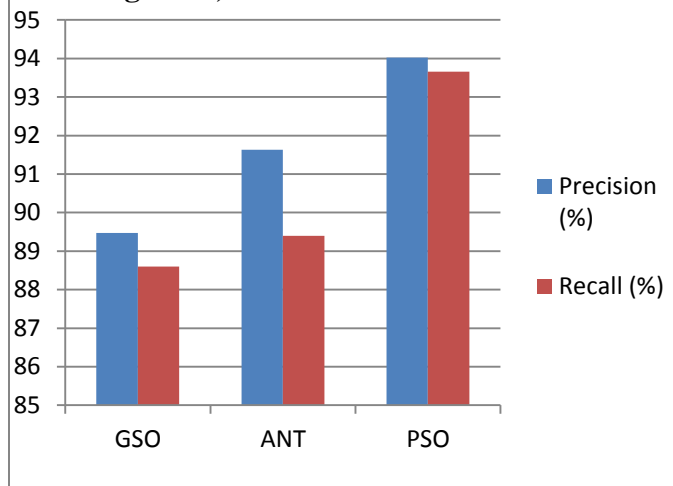


Fig. 3. Shows performance of a data set for 2000 image counts of data and rate of precision and recall on the base of GSO, ANT and PSO method. The above graph shows the comparative result analysis of image data sets of 2000 images with three different methods and calculates the parameter value of precision and recall. The variation of results in precision and recall on the basis of applying three methods found that the result of PSO is better than GSO and ANT

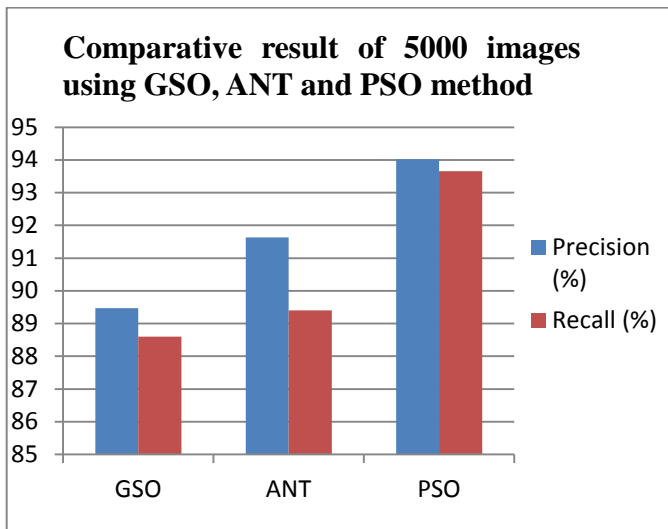


Fig. 4. Shows performance of data set for 5000 image counts of data and rate of precision and recall on the base of GSO, ANT and PSO method. The above graph shows the comparative result analysis of image data sets of 5000 images with three different methods and calculates the parameter value of precision and recall. The variation of result in precision and recall on the basis of applying three methods found that the result of PSO is better than GSO and ANT

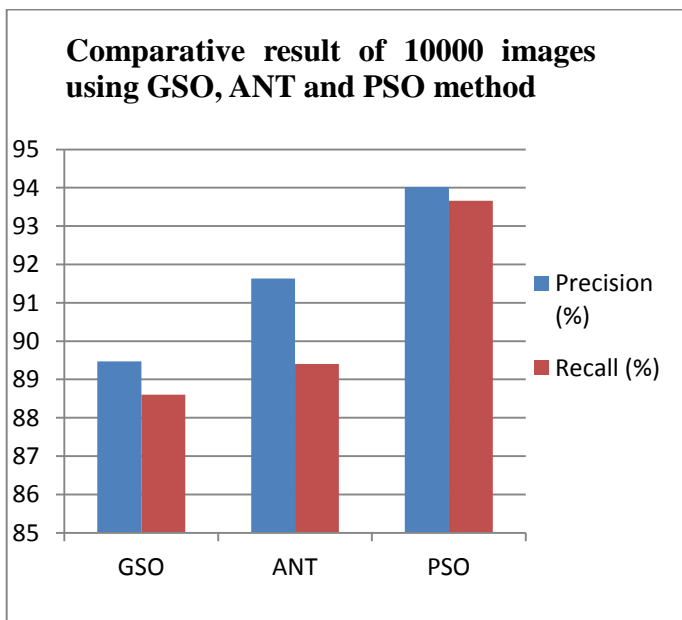


Fig. 5. Shows performance of a data set for 10000 image counts of data and rate of precision and recall on the base of GSO, ANT and PSO method. The above graph shows the comparative result analysis of image data sets of 10000 images with three different methods and calculates the parameter value of precision and recall. The variation of results in precision and recall on the basis of applying three methods found that the result of PSO is better than GSO and ANT

## V. CONCLUSION AND FUTURE WORK

This paper presents the performance evaluation of feature optimization based content-based image retrieval. In content-based image retrieval, the partial feature extraction technique is used. These partial feature extraction techniques are very efficient methods for feature extraction. For the optimization of feature, three algorithms namely ACO, GSO, and PSO were used. The diversity of particle swarm optimization is very high and due to this the optimization performance is better than glowworm and ACO algorithm. Our empirical evaluation of results shows that particle swarm optimization has better precision and recall value and is a bit different than ACO and glow worm algorithm. The ACO faced problem with feature discontinuity and does not get satisfactory results. The glowworm algorithm uses limited set of features and produces better results, but all features and their sub-sets suffered a problem of precision and recall of content-based image retrieval. In future design, feature-based fitness constraints function for ACO and glowworm algorithm will be addressed.

## REFERENCES

- [1] Liang-Hua Chen, Yao-Ling Hung, Li-Yun Wang (2012) An Integrated Approach to Image Retrieval. IEEE, Pp 695-700.
- [2] Dengsheng Zhang, GuojunLu (2004) Review of shape representation and description techniques. Elsevier, Pattern Recognition 37 Pp1-19.
- [3] Minh N. Do, MartinVetterli, (2003) The Finite Ridgelet Transform for Image Representation. IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 12.
- [4] EsmatRashedi, HosseinNezamabadi-pour (2012)Improving the Precision of CBIR Systems by Feature Selection Using Binary Gravitational Search Algorithm.The 16th CSI International Symposium on Artificial Intelligence and Signal Processing, IEEE Pp 39-43.
- [5] Y.L. Wu, C.Y. Tang, M.K. Hor, P.F. Wu (2011)Feature selection using genetic algorithm and cluster validation Expert Systems with Applications, Vol-38, Pp 2727-2732.
- [6] V.SelviDr.R.Umarani (2010) Comparative Analysis of Ant Colony and Particle Swarm Optimization Techniques International Journal of Computer Applications Volume 5- No.4.
- [7] MattiaBroilo, Francesco G. B. De Natale (2010) A Stochastic Approach to Image Retrieval Using Relevance Feedback and Particle Swarm Optimization, IEEE TRANSACTIONS ON MULTIMEDIA, VOL-12, No.4Pp 267-277.
- [8] Ayesha Salahuddin, AlinaNaqvi, KainatMujtaba, JunaidAkhtar (2012) Content based Video Retrieval using Particle Swarm Optimization 10th International Conference on Frontiers of Information Technology, IEEE, Pp 79-83.
- [9] S.W. Lin, K.C. Ying, S. Chen, Z.J. Lee (2008) Particle swarm optimization for parameter determination and feature selection of support vector machines Expert systems with applications” Elsevier , Pp 1817-1824.
- [10] Kirti Jain, Dr.Sarita Singh and Dr.Gulab Singh” Partial Feature Based Ensemble of Support Vector Machine for Content based Image Retrieval” in International Journal of Innovative Research in Computer and Communication Engineering 2013, Pp622-625.
- [11] TaoChen ,Huifang Deng and Chunru Dong” An Image Retrieval Method Based on Multi-instance Learning and PerturbativeGlowworm Swarm Optimization” in Journal of Information & Computational Science July 2013.

# Application of Artificial Neural Networks for Predicting Generated Wind Power

Vijendra Singh

Department of Computer Science and Engineering  
The Northcap University  
Gurgaon, India

**Abstract**—This paper addresses design and development of an artificial neural network based system for prediction of wind energy produced by wind turbines. Now in the last decade, renewable energy emerged as an additional alternative source for electrical power generation. We need to assess wind power generation capacity by wind turbines because of its non-exhaustible nature. The power generation by electric wind turbines depends on the speed of wind, flow direction, fluctuations, density of air, generator hours, seasons of an area, and wind turbine position. During a particular season, wind power generation access can be increased. In such a case, wind energy generation prediction is crucial for transmission of generated wind energy to a power grid system. It is advisable for the wind power generation industry to predict wind power capacity to diagnose it.

The present paper proposes an effort to apply artificial neural network technique for measurement of the wind energy generation capacity by wind farms in Harshnath, Sikar, Rajasthan, India.

**Keywords**—wind; neural network; wind power forecasting

## I. INTRODUCTION

Wind is a pollution-free renewable energy source. It is the only natural source of energy available everywhere. The consumption of electricity is rapidly increasing in the world with growing economies. India is a developing country and its economy is growing in all areas [1]. The limited power generation capacity of hydro and coal prompts us to find other energy sources. Wind energy is an alternative source but its potential regions have not been identified completely in detail [2]. The wind energy generation capacity depends on variability of its speed. It is necessary to investigate and collect data regarding availability of wind energy in all regions of the country for preparation of an inventory. We can predict the wind energy generation power by using soft computing techniques. Due to the variable nature of wind speed, wind energy forecasting is a crucially important factor. The forecasting of wind energy generation capacity involves several modeling techniques that use weather and past wind energy generation data [3, 4]. In order to achieve the highest forecasting accuracy, the energy forecasting techniques should take appropriate data to show future trends. Researchers and scientists have developed a number of energy estimation and prediction techniques for wind energy produced by wind farms [5, 6, 12]. In their works, artificial intelligence methods such as neural networks [5] and fuzzy logic are found efficient

and accurate in comparison to traditional statistical methods [3, 4, 10-17].

In this paper, a wind power forecasting method is presented based on the neural network methodology. The input parameters to this proposed model are wind speed, relative humidity, and generation hours. The output parameter is the generated wind energy by wind turbine. The wind energy relative accuracy is measured in terms of Mean Squared Error (MSE) and the Mean Absolute Error (MAE).

## II. WIND ENERGY STATUS IN RAJASTHAN

USA ranks first amongst the wind energy generation countries in the world. The wind energy production capacity of India is 102,788 MW at 50 meters above the ground level [1, 7]. India holds the fifth position in the generation of wind energy after USA, China, Germany, and Spain. The state of Rajasthan is in the northwestern part of India. It is located between 23° and 30° and 30° and 11' on the northern latitude and 69° and 29' and 78° and 17' on the east longitude. It covers 132,140 square miles of total geographical area and the largest state of India. The transmission power grids of Rajasthan connect to five neighbor states through 400-KV lines. It has a capacity of 3593.745 MW power generations by nuclear, hydropower, and thermal [8]. The location of the Aravali mountain range in Rajasthan along the north east to south west is favorable for wind energy generation. A total wind power generation capacity of 1639 MW is achieved by India including, Tamil Nadu- 895MW, Maharashtra- 399MW, Gujarat- 166 MW, Andhra Pradesh 92.6 MW, Karnataka- 93.00 MW, and Rajasthan - 60 MW. The state of Rajasthan has a gross potential of 5400 MW and technical potential of 845 MW of wind energy [7, 9]. Rajasthan has experienced less attraction for private sector compared to other states in India to generate wind power. The Harshnath mountain region in the Sikar district of Rajasthan is a unique location for positioning all wind turbines on a single place with installed capacity of 12MW. The wind turbines are scattered along the Harshnath from Goreia village to the Jeenmata temple. The winds of south west monsoon blow at an average speed of 11.5-12.5 m/s from June to August. The wind blows at an average speed of 5.5- 6.5 m/s yearly in this area.

## III. ANALYSIS OF AVAILABLE WIND POWER

The wind turbine generates wind power as kinetic energy. The density of air and speed directly influence the generation

power of wind capacity [1]. The power generation is calculated as

$$P_{wind} = \frac{1}{2} \sigma A V^3, \quad (1)$$

The wind power  $P_{wind}$  is proportional to  $\sigma$  air density in  $kg/m^3$ ,  $A$  denotes the swept area of wind turbine in  $m^2$ , and  $V$  is up stream wind speed in meter per second, The turbine power output influences on the wind capacity because power is cubic proportional to wind speed in watts. The air density variation can be measured diversely during entire year on various locations. It has lesser diversity in comparison to wind variation. The wind turbine power production capacity is proportional to wind speed during all the seasons of a year [10]. The design specification of the wind turbine controls the power generation capacity such as length of blade and cut in speed, the lowest wind speed to start the turbine for generating useful power, rated wind speed on which the machine generates the rated wind generation, and cut-out speed that allows production of maximum power [11]. In general, the cut-in speed of air is accepted by the wind turbine between 2.5  $m/s$  to 3.5  $m/s$  and the cut-out speed is in the range of 20 to 25  $m/s$ . The wind farm generates rated power when wind speed received by the machine is higher than rated wind speed (approximately 14  $m/s$ ). The average power generation capacity of wind farms fluctuated between 0.25 at 7.60  $m/s$  to 0.044 at 2.85  $m/s$  in monthly basis. It clearly shows the variation in wind energy and an area of research for prediction of wind energy.

#### IV. INPUT PARAMETERS FOR WIND POWER GENERATION

The setting of input parameters in neural network system is a crucial task for prediction of wind energy. Different parameter values might yield very different power generation results. A good setting of parameters for a neural network system may give best solution within a reasonable time period.

The input parameters of wind energy prediction system are wind energy generated, generation hours, temperature, average pressure, and relative humidity. The average-monthly input parameters of neural network based forecasting system are relative humidity, wind speed, and generation hours. All the input parameters are collected for wind power calculation and analysed to define relationship between the input and output by using correlation method. The output parameter of forecasting is highly dependent on two input parameters, average wind speed and the generator hours with a high correlation values. The relative humidity has a weak affect on the output parameter with a moderate correlation. The two input parameters called average temperature and average pressure have not been examined during modeling neural networks because both parameters have low variations. The input parameter relative humidity of wind depends on the amount of water vapor in the air and it has an effect on the air density [1]. The air density is inversely proportional to relative humidity. The air density is calculated as

$$D_{Air\ density} = d \left( \frac{273.15}{T} \right) \times \left[ \frac{B - 0.3783e}{760} \right]. \quad (2)$$

Here  $d$  represents the density of the dry air at atmospheric temperature (25° C) and standard pressure (100 kPa) ( $d = 1.168\ kg/m^3$ ).

The air density is inversely proportional to  $T$  absolute temperature (Kelvin) and directly proportional to the barometric pressure  $B$  (torr). The symbol  $e$  is vapor pressure of moist air in torr. The proposed neural network wind power forecasting system takes three input parameters, wind speed, relative humidity and generation hours.

TABLE I. WIND GENERATORS CAPACITY

Wind generator	WG1	WG2	WG3	WG4	WG5	WG6
Capacity (MW)	0.60	0.60	0.60	0.60	0.60	0.60

#### V. NEURAL NETWORK FORECASTING SYSTEM FOR WIND ENERGY

##### A. Biological Neuron

Human neuron system, a biological neuron receives inputs from all parts of body, combines the inputs, performs a nonlinear operation and gives the final output result. Figure 1 depicts the correlation of these parts.

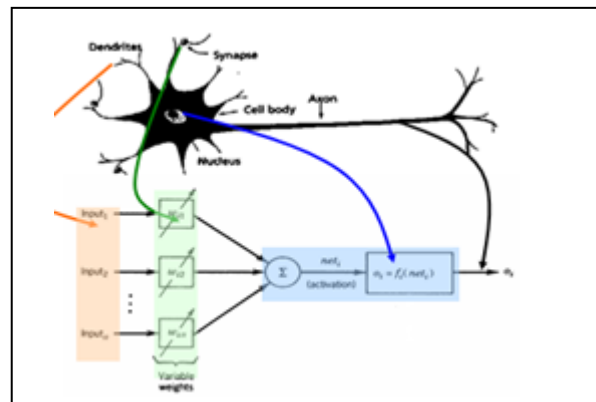


Fig. 1. A correlation between biological neuron and ANN

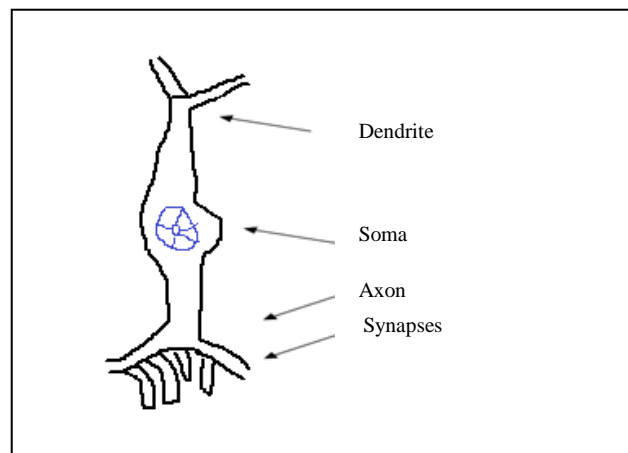


Fig. 2. A Neuron Components



In our body system, there are many variations of different neurons. A neuron consists of four components known as soma, dendrites, axon, and synapses as shown in Figure 2. The dendrite is very thin like hair and is an extension of the soma and works as input channels [5]. It receives its input signals through the synapses of other neurons of our human body. The soma processes these incoming signals and sends to other neurons through the axon and the synapses.

**B. Artificial Neural Network**

An Artificial Neural Network (ANN) is a type of model that was first abstracted by McCulloch and Pitts [5]. ANNs were designed on biological neurons system, and how neurons interact with each other in the brain. Kilogram stated that during the past years, there has been a substantial increase in the interest on the ANNs. The scientists and researchers have designed and developed ANN systems for application in various areas of social sciences, physical science, and engineering, medicine and many other areas [5]. The neural network based system needs to train on an input data set known as training data set. Thereafter, the system is tested by using test data set and this process is known as testing. Its results are validated before prediction of data sets [12]. The ANN can be used to solve complex applications that are not possible by the traditional approaches and difficult to model analytically [14, 15]. The basic processing unit of an artificial neural network is a neuron. It takes input, processes on input and gives output. The ANN can be designed using either feed forward or feedback approach. There are three types of layers, input layer, hidden layer, and output layer in an ANN [13, 16]. The ANN neuron model is depicted in Figure 3. Each input  $i$  to neuron has an associated weight  $w_i$ . A neuron  $j$  can be mathematically represented by the following expressions [5].

$$u_j = \sum w_{ji} x_i \tag{3}$$

$$y_j = f(u_j - \theta_j) \tag{4}$$

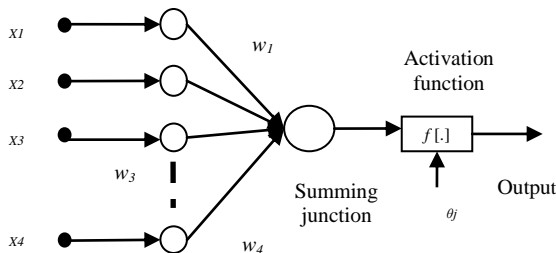


Fig. 3. Neuron model

**C. ANN Architecture for Wind Energy**

The proposed neural network for wind power forecasting model is a feed forward network model with supervised learning using back propagation algorithm [5]. The overall structure of wind power forecasting model is 3-4-1 as shown in Figure 4. The input layer receives three inputs, average wind speed, average relative humidity, and average generation hours per month, processed by three processing units. We set bias value equal to zero during calculation in activation functions. We have adopted the logarithmic sigmoid function as activation function at the second layer and linear activation function at the output layer. The weights at each of

intermediate and output layer are adjusted in the training mode until the errors are detected within the pre specified range. The output parameter is the wind energy generation by the wind farms. The proposed model is implemented using MATLAB to forecast wind energy by wind generators as shown in Figure 4. The wind power forecasting model has trained in batch mode [19]. The batch training is typically done in train mode and inputs are dynamically converted into concurrent vectors. A new vector generates to show how close the input is to the target vector. The input parameters values are scaled within the range of 0 -1 during normalization.

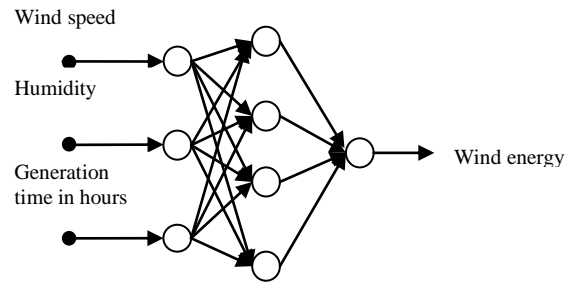


Fig. 4. ANN Architecture for wind energy forecasting

**D. Training performance and accuracy of the Prediction**

In the proposed wind power forecasting system to determine quantitatively evaluation of the model, two error detection methods are used. The accuracy of model can be determined by mean squared error (MSE) as shown in (6) and mean absolute error (MAE) as shown in (7) [1, 5]. The MAE is a quantity method used to measure how forecasts are close to each other. In this model, we trained the neural network in range from 10-min, to 60-min time delay modes.  $y_t$  is the  $t$ th observation of the actual wind power measurement and  $F_t$  is predicted wind power for same period of time  $t$ . The actual wind power production by turbine is presented in Table 2. The error  $e_t$  can be calculated as in (5).

$$\varepsilon_t = \psi_t - \Phi_t \tag{5}$$

$$MSE = \frac{1}{n} \sum_{i=1}^n e_i^2 \tag{6}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |e_i| \tag{7}$$

where  $n$  is length of vector for wind power forecast data. When the MSE is reduced in a gradual way and its became stable, than wind power forecast model gives satisfactory results on training and testing data sets. Figure 5 shows the training accuracy of wind power forecast model. The performance of the forecasting network is tested by using actual wind energy output. The accuracy is assessed using two methods. In the first method, the forecasted wind power output is compared with the actual wind power production. Figure 6 shows accuracy of the forecasted wind power generation per hour. The average percentage error analysis is 7. In the second method, RMSE MAE are measured and analysed. The error estimation using MAE and MSE statistics of the forecasting results are shown in Table 2. The goal set of MSE is 0.0001. The proposed model has been run for 1200 iterations but MSE of training data set become static at 0.0070

after 300 iterations. Figure 6 shows predicted and actual wind power representation. The predicted and actual wind power results closeness shows the accuracy of forecasting model.

TABLE II. ERROR ANALYSIS

Input (MW)	MAE	MSE
10 min	4.82	8.20
20 min	5.45	9.15
30 min	6.55	10.30
40 min	7.15	11.25
50 min	8.25	12.00
60 min	9.35	13.50

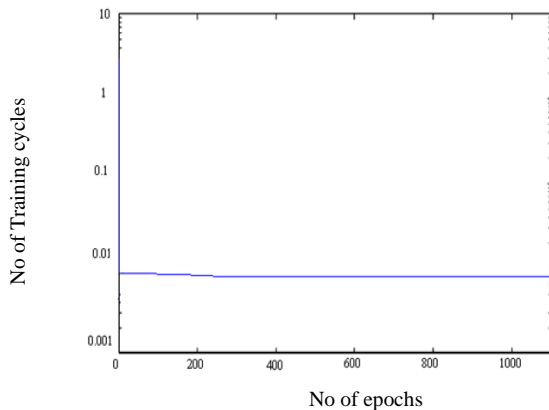


Fig. 5. Training performance of the neural network

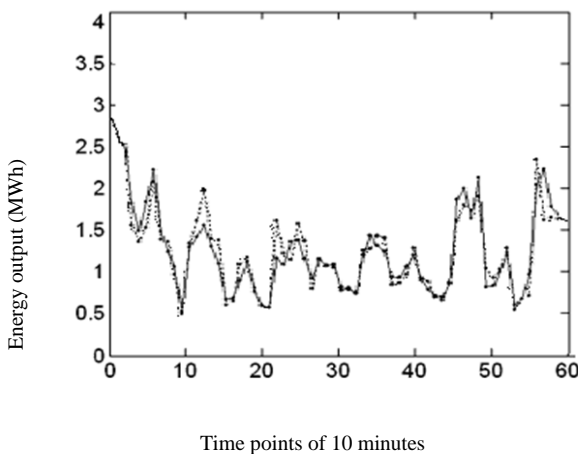


Fig. 6. Measured and predicted wind power

## VI. CONCLUSION

This paper presents a wind power forecasting model based on feed forward neural network architecture. The back propagation algorithm was applied to trained proposed model using supervised learning technique. The data are divided in two sets, training data and test data for testing accuracy of proposed model. The input parameters, average wind speed, average relative humidity, and generation hours are collected from various sources.

The performance of model is measured by using two error-detection methods, MSE and MAE. The proposed model was run at different iteration both on training and testing data sets. The MSE becomes stable at 0.0070 at 300 iterations. The proposed model wind power prediction shown in Figure 6 indicates its capability of forecasting. It proves to be an efficient and valuable tool for the estimation of wind energy for wind generators.

## REFERENCES

- [1] Mabel MC, Fernandez E. Analysis of wind power generation and prediction using ANN: a case study. *Renewable Energy*, Vol. 33(5):986-92, 2008.
- [2] F. Manwell, J. G. McGowan, and A. L. Rogers, *Wind energy explained: theory, design and application*. New York: Wiley, 2002.
- [3] Shuhui Li, Donald C. Wunsch, Edgar A. O'Hair and Michael G. Giesselmann, Using neural networks to estimate wind turbine power generation, *IEEE Trans. on Energy Conversion*, vol. 16, pp.267-282, 2001.
- [4] Andrew Kusiak, Haiyang Zheng and Zhe Song, Short-term prediction of wind farm power: a data mining approach, *IEEE Trans on Energy Conversion*, vol.24, no.1, pp.125-135, 2009.
- [5] S. Haykin, *Neural networks, a comprehensive foundation*, New Jersey, 2nd edition, 1999.
- [6] D. A. Bechrakis and P. D. Sparis. Correlation of wind speed between neighboring measuring stations, *IEEE Trans. Energy Convers.*, vol. 19(2), pp. 400-406, 2004.
- [7] Wind power in India, <http://www.windpowerindia.com/index.php> [accessed on 01.02.16].
- [8] Ministry of new and renewable energy, <http://www.mnre.gov.in/> [accessed on 01.01.16]
- [9] S. Mani, T. Dhingra. Policies to accelerate the growth of offshore wind energy sector in India. *Renewable and Sustainable Energy Reviews*, vol 24 (c), pp. 473-482, 2013.
- [10] Lei, M., Shiyang, L., Chuanwen, J., Hongling, L., and Yan, Z. A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews*, vol 13(4), 915-920, 2009.
- [11] Kaldellis, John K., and D. Zafirakis. The wind energy (r) evolution: A short review of a long history. *Renewable Energy*, vol 36, no. 1887-1901, 2011.
- [12] Indu R. Pillai, Rangan Banerjee, *Renewable energy in India: Status and potential*, *Energy*, vol 34(8), pp. 970-980, August 2009.
- [13] T. G. Barbounis, J. B. Theocharis, M. C. Alexiadis, and P. S. Dokopoulos. Long-term wind speed and power forecasting using local recurrent neural network models, *IEEE Trans. Energy Convers.*, vol. 21, no. 1, pp. 273- 284, 2006.
- [14] Bilgili, M., Sahin, B., & Yasar, A. Application of artificial neural networks for the wind speed prediction of target station using reference stations data. *Renewable Energy*, vol 32(14), 2350-2360, 2007.
- [15] P. Flores, A. Tapia, and G. Tapia. Application of a control algorithm for wind speed prediction and active power generation, *Renewable Energy*, vol. 30, pp. 523-536, 2005.
- [16] Fadare, D. A. The application of artificial neural networks to mapping of wind speed profile for energy application in nigeria. *Applied Energy*, vol 87(3), 934-942, 2010.
- [17] Pourmousavi Kani, S. A., and Ardehali, M. M. Very short-term wind speed prediction: a new artificial neural network-markov chain model. *Energy Conversion and Management*, vol 52(1), 738-745, 2011.
- [18] Mahamad, Nurhashimah, and Muhamad Kamal B. Mohammed Amin. Power prediction analysis using artificial neural network in ms excel. 7th International Conference on Latest Trends in Renewable Energy and Environmental Informatics, Malaysia, pp 54-58, 2013.
- [19] Karabacak, K., and Cetin, N. Artificial neural networks for controlling wind-pv power systems: a review. *Renewable and Sustainable Energy Reviews*, vol 29, 804-827, 2014.

# Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data

Suhail Sami Owais  
Dept. of Comoputer Sciecn  
Applied Science Private University  
Amman, Jordan

Nada Sael Hussein  
Master of Computer Science  
Amman, Jordan

**Abstract**—There is an exponential growth in the amount of data from different fields around the world, and this is known as Big Data. It needs more data management, analysis, and accessibility. This leads to an increase in the number of systems around the world that will manage and manipulate the data in different places at any time. Big Data is a systematically analysed data that depends on the existence of complex processes, devices, and resources. Data are no longer stored in traditional database storage types or on such forms like database, which only on structured data are limited, but surpassed them to the unstructured or semi-structured data. Thus, Big Data has several characteristics and specific properties proportionate with the size of the data with the enormous and rapid development in all business and areas of life. In this work, we study the relationship between the characteristics of Big Data and extract some categories from them. From this, we conclude that there are five categories, and these categories are related to each other.

**Keywords**—Big Data; Characteristics; Categories; Management; Analysis; Anywhere and Anytime

## I. INTRODUCTION

Big Data refers to technologies and initiatives that involve data that is too diverse, rapidly changing or massive for conventional technologies, skills and infrastructure to address efficiently. In other words, the volume, velocity, or variety of data is too great [1]. Big Data requires new technologies with a spatial architecture so that it becomes possible to extract value from it by capturing and analysis process[2]. Due to such large size of data (rising up information yields an increase in the amount of data), it becomes very difficult to perform effectively and to analyse using the existing or the known traditional techniques [3].

Big Data due to its various properties like volume, velocity, variety, variability, value, and complexity puts forth many challenges [4]. Since Big Data is an upcoming technology in the market which can provide many benefits to the business organizations, it becomes necessary that various challenges and issues associated in investigating and adapting to this technology are brought into light.

The main data categories are: the Traditional Database category and the Big Data.

### A. Traditional Database:

A database is a set of relevant data by data, which means known facts that can be recorded and have implicit meaning. For example, names; telephone numbers, and addresses of the people you know. You have recorded these data in an indexed

address book or stored them on a hard drive, using a personal computer and software such as Microsoft Access or Excel. The collection of the related data with an implicit meaning is known as a database [5].

### B. Big Data:

Big Data is the term used to describe huge amount of structured and unstructured data which is large. It was very difficult to process the data in the Big Data using the traditional databases and software technologies [6]. Also Big Data is companies who had to query loosely structured very large distributed data [7].

This research consists of five sections. The first one is an introduction about Big Data and its category. The second section discusses how Big Data becomes growth and from where it comes to be called as Big Data. It also demonstrates how the authors categorize the characteristics of Big Data into five categories and discuss them. The third section explains what are the benefits of Big Data. The fourth section demonstrates that Big Data can be found anywhere, anytime, and in anyplace. Finally, the last section offers conclusion and the future work.

## II. BIG DATA

An exponential growth in the amount of data from all different types of data becomes enormous data known as Big Data. Thus, Big Data is a massive set of structured and unstructured and semi-structured data [8]. It is hard to manage the data in the Big Data using any of the traditional applications [9]. Big Data consists of very large datasets that will be processed by any traditional database system tools [6].

Collecting data in Big Data is done not only from the traditional corporate database only, but includes other sources as well:

1) *Data collected from Sensors*: Sensors are becoming an increasingly important component of the information stored and processed by many businesses. For this a lot of data were collected from different fields of sensors. This includes data from Fixed-Sensors such as home automation; traffic sensors; traffic-webcam sensors; scientific sensors; security-monitoring videos or images and from weather-pollution sensors as well as data from Mobile-Sensors such as data from GPS Sensors; mobile phone location, satellite images

2) *Data collected from Machines*: Where there are a lot of data collected form machines in which it is collected from

sensor data, it is known as complex data. For example: Video from security cameras, Recording voice form microphone, Mail to Mail Log Files, Satellite Imaging and Bio-Informatics.

3) *Data collected from Human:* Several data were collected from humans. This included data from human enterprise contents and from external sources and from Documents, E-mail, Web Logs and social networks like Facebook; LinkedIn; Twitter; Instagram; Flickr, and Picasa.

4) *Data collected from Business Process:* Data were collected from industry, production, refining, distribution, and from Marketing. For example, data produced by public agencies like that from medical records, data produced by businesses such as commercial transactions, and from banking records and from E-commerce and credit cards.

Big Data has many characteristics or properties mentioned by nV's characteristics [8]. Set of V's characteristics of the Big Data were collected from different researchers' publications to have Nine V's characteristics (9V's characteristics). These 9V's characteristics are: (Veracity, Variety, Velocity, Volume, Validity, Variability, Volatility, Visualization and Value).

We categorize these characteristics of the Big Data into five categories CPIVW. They are: Collecting Data, Processing Data, Integrity Data, Visualization Data and Worth of Data.

Figure 1 demonstrates the Big Data CPIVW categories and their characteristics. The CPIVW five categories will be discussed in the next section.

### 2.1 Five Categories CPIVW of the Big Data and their 9 V's Characteristics:

Because of the relationship between some of the characteristics, they are all classified into different categories. The characteristics of the Big Data they clustered into groups as categories with respect to the relationship between them. Thus, we extract five categories CPIVW (Collecting Data, Processing Data, Integrity Data, Visualization Data and Worth of Data) from these 9V's characteristics of the Big Data.

The extracted five categories CPIVW (Collecting Data, Processing Data, Integrity Data, Visualization Data and Worth of Data) are discussed below:

#### 1) *Collecting Data:*

Data is collected from different resources and different types to have Big Data as a massive set of structured data and unstructured data and semi-structured data. For this veracity and variety characteristics for the Big Data have such relationship between them. So, they grouped together to form collected data category.

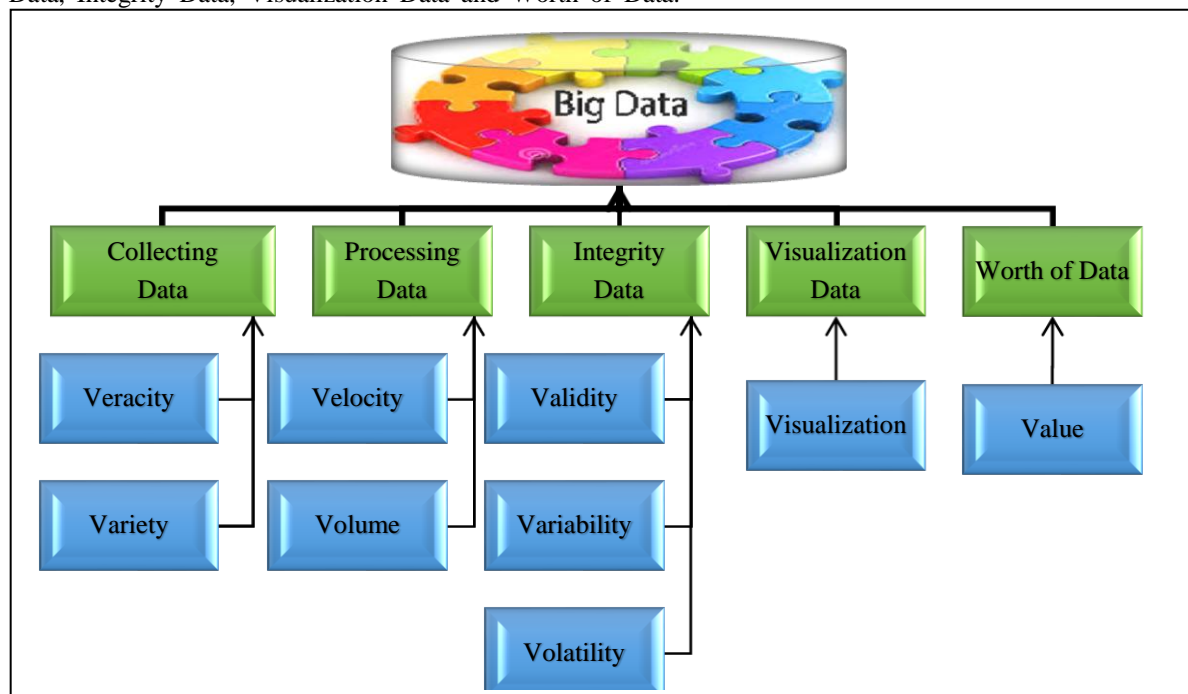


Fig. 1. Five Categories CPIVW of the Big Data with their 9 V's Characteristics

- **Veracity:** Big Data veracity refers to the biases, noise, and abnormality in data. The data that is stored, and mined meaningful to the problem that is being analysed, in addition the developers ask the question "Is the data that is being stored, and mined meaningful to the problem being analyzed or not?" [6]. The veracity not only talks about the quality of data, but when the users begin using the Big Data, they also become truly engaged and are more willing to invest in efforts to clean up data ideally at the source [10].
  - **Variety:** Structured, semi-structured, and unstructured data besides text and more data types have emerged, such as record, log, audio, and hybrid data [11]. Currently, data comes in all types of formats such as emails, video, audio, transactions, and images. In other words, the variety is structure, semi-structure, and unstructured data in same place [12].
- 2) *Processing Data:*

Processing Data comes from grouping the two main V's characteristic of Big Data, velocity and volume, because there exists such a relationship between them. The processing data category talks about speed of processes on the data fit with size of Big Data. The data in processing data category passes through many of the processes and is processed on demand. Here is a description of the two Vs, velocity and volume:

- **Velocity:** The created information at faster pace than before, in which the different channels of Big Data increase the output content. This property means how fast the data is to be produced and processed to meet the demand [6].
- **Volume:** It is predicted that the data volume worldwide will reach 40 ZB by 2020 [11]. Storing different type of data from social network is possible in Big Data storage devices. For this, the amount of data is known as volume of data, where the amount of data continues to explode. Thus, to improve the company's for archiving, and tiered data importance strategies to accommodate the new volumes? There are many factors which contribute to increasing the volume streaming data and data collected from sensors and other resources. The large volume of Big Data is the primary goal of consumers to optimize future results [6]. From this, the main goal from the large volume of Big Data is to make it useful for users and consumers and optimize future results. Thus, such a programming model (MapReduce) that associated implementation for processing and generating large data sets [13].

### 3) Integrity Data:

Validity, variability, and volatility are the three Vs grouped together to categorize the data integrity. Data integrity refers to the accuracy and consistency of data stored in Big Data, and talks about the truth of data and guarantees the data as correct and not manipulated. Integrity data also ensures the quality of the data in the Big Data.

- **Validity:** As such, Big Data veracity is a matter of validity, meaning that the data is correct and accurate for the intended use. Clearly valid data is the key for making the right decisions [6]. Data validation is one that certifies uncorrupted transmission of data.
- **Variability:** Along with the velocity, the data flows may be highly inconsistent with periodic peaks, daily, seasonal, and event-triggered peak data loads can be challenging to manage, especially with unstructured data involved [14].
- **Volatility:** When we talk about volatility of Big Data, we can easily recall the retention policy of structured data that we implement every day in our businesses [15]. Once retention period expires, we can easily destroy it. For example: an online ecommerce company may not want to keep a one year customer purchase history. Because after one year and default warranty on their product expires so there is no possibility of these data restore ever [15].

### 4) Visualization Data:

Visualization data refers to a way to explore and understand your data, in the same way that the human brain processes information. Visualization Data category contains only the visualization characteristic for Big Data, thus the data is easy to read and analyse from complex graphs. Visualizing the data in the Big Data leads the people to understand the meaning of different data values faster when they are displayed in charts and graphs rather than reading about them from reports.

- **Visualization:** It is the hard part of Big Data which makes all that huge amount of data comprehensible and easy to understand and read. With the right analysis and visualizations, raw data can be put to use, otherwise raw it remains essentially useless [16]. Visualization means complex graphs that can include several variables of data while still remaining understandable and readable [17].

### 5) Worth of Data

Big Data characteristic value is the worth of data category. Worth of data talks about the cost and management of data in Big Data. Small data can have more value than a corresponding Big Data collection, and the Big Data has a high economic value. The collected data comes with many noises. While collecting data, it may contain some noises, and thus it must be filtered from any noise to help the user to analyse and take decision.

- **Value:** It has a low-value density as a result of extracting value from massive data. Useful data needs to be extracted from any data type and from a huge amount of data [8]. For this we must look for true value of data, in which data value must exceed its cost or management. Thus, attention must be paid to the investment of storage for data. Storage may be cost effective and relatively cheaper at the time of purchase but such underinvestment may damage highly valuable data [18]. For example storing clinical trial data for new drug on cheap and unreliable storage may save money today but can put data on risk tomorrow [19].

### 2.2 Big Data CPIVW Categories Hierarchy

After CPIVW has been clarified, the 9V's characteristics of Big Data were grouped in clusters to get the five categories CPIVW. And from this we present them in a hierarchy model as shown in Figure 2. Here it is shown that the five categories CPIVW have some dependency to get the Big Data in such manner.

### III. SOME BENEFITS FROM THE BIG DATA

Distributing data around the world via Internet and the increasing amount of data tends to become vast and huge data. Thus, a lot of data can be created on a single day itself, and from where these data's collected from where. It comes from around the world and it may be more than 2.5 quintillion byte from different resources. This yields to begin most of the companies with different sizes to gain the benefits of data technology and analytics. While if your company was not up to speed so far? For this, at hand there are several ways that Big Data can help with several benefits your business. But, may or may not there is/are such obstacles that can thwart the Big Data plans.



In Information technology (IT), the executives continually evaluate the technology trends that will impact their business. Some simply deploy technology to promote the goals stipulated in business plans. Others take on the role of chief innovation, creativity and renewal officer and introduce different models of using existing data to generate new revenue and gain insight into who clients are and what they want.

Any IT organization considering a Big Data initiative should consider these five major points [20], which will bring clarity as well as revenue and benefits to a company.

Several benefits from the Big Data that may help the business are discussed as follows:

- Management Data in Big Data using an intelligent tools will be better which enable many use cases. While data may be founded in different formats that it has been successfully mined for specific purposes.
- Internet benefits. The benefit use of cloud storage (storing of data online in the cloud or in "the Internet"), and the benefit use from cloud services providers (servers, storage and applications through the Internet). Both benefits help the business from storage and the cloud computing power. When files are stored in the cloud, they can be accessed at anytime from anywhere via Internet and accessed with remote backups of data.
- Visualizing data from Big Data using business intelligence software is presented in a simple way to read and analyse. This software must be able to provide the processing engines that allow the end users to query and manipulate information quickly evenly in real time in some cases.
- Capabilities will evolve the Business data analysis methods for data from the Big Data. In which a structured data; unstructured data and semi structured data are grouping as with different types of data to form the Big Data that forms the data in such business. Some examples are text file and audio and video files. Thus, an intelligent tool needs to be used for recognizing a specific pattern based on such criteria. A natural language processing tool can prove vital to text mining, sentiment analysis, and entity recognition efforts.

#### IV. BIG DATA ANYWHERE-ANYTIME-ANYPLACE

Big Data yields to explore such technology that will permit and control the enormous data founded around the world. These data can be accessed via Internet from anywhere at any time and from any platform. One of the most important technologies were founded is called Cloud Computing. Big Data does not refer only to the huge amount of data, but also it refers to the computational term in a large number of fields either economic; business; medical; researches and any other fields. Because of rising in the amount of data in the Big Data which yields to increase the demands for accessing the data resources from anywhere at any time from anyplace.

Several challenges must be noticed here instead of using just single computer and accessing the data from only one storage device. Thus a software's applications must be founded

and abled to execute on anywhere over the networked devices. Via internet where each device in the networked devices may join or leave the shared data scope at anytime from anywhere at any place. Thus, for processing data in the Big Data the founded software may use concurrency to take advantage of multiple foundations.

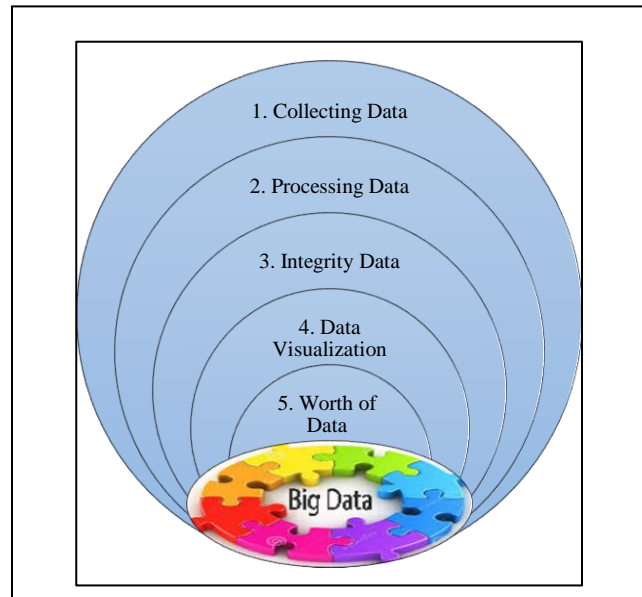


Fig. 2. Five Categories CPIVW Hierarchy for the Big Data

#### V. CONCLUSION AND FUTURE WORK

With the increasing amount of data, it has become difficult to be handled using such traditional tools, as that used for data bases. Big Data appears to coincide with the tremendous development in all areas of the life that led to the aggravation of the size and diversity of data sources such as data collected from Sensors; Machines; Human; and from Business Process. Data from different sources cause Big Data to be a massive set of structured and unstructured and semi-structured data. For this, it becomes too hard to manage the data in the Big Data by using a traditional applications.

Big Data has 9V's characteristics (Veracity, Variety, Velocity, Volume, Validity, Variability, Volatility, Visualization and Value). The 9V's characteristics were studied and taken into consideration when any organization need to move from traditional use of systems to use data in the Big Data.

The relationships between Big Data characteristics are studied and we extract five categories by assortment the 9V's characteristics of the Big Data, also the five categories known as CPIVW and they are (Collecting Data, Processing Data, Integrity Data, Visualization Data and Worth of Data).

Last but not least, it must be said that Big Data needs to be an environment fit with the size and data types. This environment is a cloud computing environment in which Big Data emerged and the type we will look at the next paper to the mechanism of Big Data work and how they are managed and analyzed within the cloud computing environment and what types of that fit better than others with large data and taking



into consideration several aspects of things, and we will look to in a timely manner.

#### ACKNOWLEDGMENT

The authors are grateful to the Applied Science Private University, Amman, Jordan, for the full financial support granted to this research project (Grant No. DRGS-2015-2016-61).

#### REFERENCES

- [1] Big Data, mongoDB White Paper, available at: <https://www.mongodb.com>, February 2015.
- [2] K. Li, H. Jiang, L. T. Yang, A. Cuzzocrea, "Big Data: Algorithms, Analytics, and Applications", 1<sup>st</sup>. edition, Taylor & Francis website, pp. 91-104, 2015.
- [3] S. Gupta, M. S. Chaudhari, "Big Data Issues and Challenges: Data analysis, storing, processing, issues, challenges and future scope", International Journal on Recent and Innovation Trends in Computing and Communication, vol. 3, Issue 2, pp. 62-67, February 2015.
- [4] L. V. Satyanarayana, "A Survey on Challenges and Advantages in Big Data", International Journal of Computer Science And Technology, vol.6, Issue 2, April-June 2015.
- [5] R. Elmasri, S. B. Navathe, "Fundamentals of database systems," 6th edition, Adeson-Wesley publishes, pp. 3-24, New York, 2011.
- [6] Vangie Beal, "Big Data", Available at: [http://www.webopedia.com/TERM/B/big\\_data.html](http://www.webopedia.com/TERM/B/big_data.html), February 2016.
- [7] A. Katal, M. Wazid, and R. H. Goudar, "Big data: issues, challenges, tools and good practices," in Proceedings of the 6th International Conference on Contemporary Computing (IC3 '13), pp. 404-409, IEEE, 2013.
- [8] Nour E. Oweis, Suhail S. Owais, Waseem George, Mona G. Suliman, and Václav Snášel. A Survey on Big Data, Mining: (Tools, Techniques, Applications and Notable Uses). Intelligent Data Analysis and Applications, Springer Cham Heidelberg New York Dordrecht London. Volume 370, 2015.
- [9] A. Syed, K. Gillela, and C. Venugopalnd, "The Future Revolution on Big Data", International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, Issue 6, June 2013.
- [10] Patricia Saporito, "2 More Big Data V's — Value And Veracity," Digitalist Magazine by SAP, Available at: <http://www.digitalistmag.com/big-data/2-more-big-data-vs-value-and-veracity-01242817>, 2014.
- [11] Full lifecycle management of Big Data White Paper, available at: [www.huawei.com](http://www.huawei.com), 2013.
- [12] Infographics, "The four V's of Bog Data," IBM Big Data & Analytics Hub, Available at: <http://www.ibmbigdatahub.com/infographic/four-vs-big-data>, February 2016.
- [13] J. Dean, S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communication of ACM-50th anniversary issue: 1958-2008, vol.51, no. 1, pp. 107-113, Jan 2008.
- [14] R. Toshiwal, K.G. Dastidar, and A. Nath, "Big Data Security Issues and Challenges," International Journal of Innovative Research in Advanced Engineering (IJIRAE), vol. 2, issue 2, February 2015.
- [15] M.A. Khan, M.Fahim Uddin, and N. Gupta, "Seven V's of Big Data," Proceedings of 2014 Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1), Bridgeport, Connecticut, USA, pp.3-5, April 2014.
- [16] V'S of Big Data, Available at: <http://www.godatafy.com/category/tech-blog/big-data/>
- [17] Mark van Rijmenam, "Why The 3V's Are Not Sufficient To Describe Big Data," Datafloq, Available at: <https://datafloq.com/read/3vs-sufficient-describe-big-data/166>.
- [18] V.S.Thiyagarajan, K.Venkatachalapathy, "Isolating Values from Big Data with the Help of Four V's", International Journal of Research in Engineering and Technology, vol.4, issue 1, January 2014.
- [19] J. Hurwitz, A. Nugent, F. Halper, and M. Kaufman, "Big data for Dummies," John Wiley & Sons, Inc. publishes, Hoboken, New Jersey, 2013.
- [20] Reda Chouffani, "5 Reasons to Move to Big Data (and 1 Reason Why It Won't Be Easy)," Serving Chief Information Officers (CIOs), Available at: <http://www.cio.com/article/2385690/big-data/5-reasons-to-move-to-big-data--and-1-reason-why-it-won-t-be-easy-.html>, 2013.

# Hybrid Solution Methodology: Heuristic-Metaheuristic-Implicit Enumeration 1-0 for the Capacitated Vehicle Routing Problem (Cvrp)

David Escobar Vargas

Faculty of Electrical Engineering  
Technological University of Pereira  
Pereira, Colombia

Ramón A. Gallego Rendón

Faculty of Electrical Engineering  
Technological University of Pereira  
Pereira, Colombia

Antonio Escobar Zuluaga

Faculty of Electrical Engineering  
Technological University of Pereira  
Pereira, Colombia

**Abstract**—The capacitated vehicle routing problem (CVRP) is a difficult combinatorial optimization problem that has been intensively studied in the last few decades. We present a hybrid methodology approach to solve this problem which incorporates an improvement stage by using a 1-0 implicit enumeration technique or Balas's method. Other distinguishing features of the methodology proposed include a specially designed route-based crossover operator for solution recombination and an effective local procedure as the mutation step. Finally, the methodology is tested with instances of the specialized literature and compared with its best-known solutions for the CVRP with homogeneous fleet, to be able to identify the efficiency of the use of the Balas's methodology in routing problems.

**Keywords**—1-0 implicit enumeration; CVRP; Operations research; Genetic algorithm; Chu-Beasley; Heuristics; Metaheuristics and exact methods

## I. INTRODUCTION

The general vehicle routing problem (VRP) is a generic name that refers to a variety of applied problems in various areas of knowledge such as transportation, supply chain, production planning, and telecommunications. The general problem is based on a group of customers that have to be served by a fleet of vehicles. They begin their tour or path in a main depot, visiting the customers assigned to the route only once, and returning to the depot where they started. The main objective is to establish the possible routes that represent the least-cost paths that will meet the customers' demands. Customers are spread geographically and associated to one or more constraints that must be met. However, the simplicity of the description is not near to the complexity of the solution search, that is classified as an NP-hard problem and ranked as one of the most interesting optimization problems in operational research. This problem has been analysed and studied extensively since its first appearance in the literature through the formulation applied to the fuel distribution by Ramser and Dantzig in 1959 [1]. From this work, several authors have focused their efforts on finding efficient techniques and models that allow to solve this problem in items (Miller et al, 1960 [2]. Christofides et al, [3][4]; Bodin's at., 1983[5]; Fisher, 1995[6]; Desrosiers et al., 1995; Powell et al. 1995; Fukasa, 2004; Gendreau, 2005a, b; Cordeau et al. 2005, 2007; Laporte, 2009 [7]), and books (Thot and Vigo, 2002 [8]; Golden et al, 2008 [9]). Looking at the state of the

problem, it is remarkable that the the large increase in algorithmic and methodical development focused on the variants of the VRP, such as the capacitated routing problem (CVRP) and the VRP with time windows (VRPTW), which highlights the techniques with better behavior for each of these variants. In this paper, the importance of the guidelines proposed in the literature were an important pillar in defining the methodology that would allow us to work on the capacitated vehicle routing problem (CVRP). One of the main objectives of this paper is seeking a technique that allows us to measure the performance of implicit enumeration 0-1 methodology or Balas's algorithm, which has not been applied to the routing problems. Therefore, our work is to define a hybrid methodology that is able to solve instances proposed in specialized literature created specifically for the CVRP.

In order to solve this problem, we propose a hybrid methodology that combines the exploration of the solution space through a search based on the evolution of a specialized population of individuals, jointly with diversity techniques in the crossover, a neighborhood search as a mutation stage and a implicit enumeration 0-1 methodology as an improvement stage. The idea is to observe the performance of an exact technique (Balas's) applied to the capacitated routing problem. The exact technique gets in charge of a sub-problem of the CVRP, as has been done in other areas [10]. The methodology called modified Genetic Algorithm of Chu-Beasley with Implicit Enumeration (GACBIE) is tested with a set of instances proposed in the specialized literature and compared to the best known solutions (BKS) for CVRP problem.

The paper is organized as follows: section II shows some related works and part of the specialized literature for the CVRP, section III shows a formal description of the capacitated vehicle routing problem followed by the methodology proposed in section IV, the pseudocode of the complete algorithm in Section V, ending with the results of tests on the proposed instances in Section VI. Finally, conclusions and recommendations from this work are illustrated.

## II. RELATED WORKS AND OTHER LITERATURE

The CVRP plays a particular role on VRP algorithmic research, for both exact and on heuristics methods. Being the most basic variant, it is a natural testbed for trying new ideas.

Its relative simplicity allows cleaner descriptions and implementations, without the additional conceptual burden necessary to handle more complex variants. Successful ideas for the CVRP are often later extended to more complex variants. For example, the classical CVRP heuristic by Clarke and Wright [11] was adapted for the VRP for many other variants, as surveyed in Penna [12].

Laporte and Nobert [13] presented an extensive survey which was entirely devoted to exact methods for the VRP and gave a complete and detailed analysis of the state of the art up to the late 1980s. Up to the end of the last decade, the most effective exact approaches for the CVRP were mainly branch and bound algorithms using basic relaxations, as the assignment problem and the shortest spanning tree. Recently, more sophisticated bounds were proposed, as those based on Lagrangian relaxations or on the additive approach, which increased the size of the problems that can be solved to optimality by branch and bound. Moreover, following the success obtained by branch and cut methods for the TSP, encouraging results were obtained by using these algorithms for the CVRP.

Since the 1980s, almost all articles proposing new heuristic and metaheuristic methods for the CVRP reported results on a subset of the instances by Christofides and Eilon [14] and Christofides et al. [15]. This classical benchmark is also exhausted, since most recent heuristics find systematically the best-known solutions on nearly all instances.

Rochat and Taillard [16] proposed an efficient local search to solve the problem with probabilistic diversification and intensification step to solve big instances of the CVRP where they report optimal solutions for almost all the instances except on a single instance with 199 customers.

In Fukasawa *et al.* [17], a branch-cut-and-price is proposed to solve all its instances with up to 100 customers, as well as instances with 121 and 135 customers. In particular, the column and cut generation algorithms of Baldacci *et al.* [18] significantly reduced the CPU time required to solve many instances, but could not solve the some larger instances. Later, some methodologies as the ones proposed on Contardo and Martinelli [19] and Røpke [20] were capable of solving bigger instances.

In a recent work, the heuristics of Vidal et al. [21] are considered one of the best available for the CVRP basically due to their superior performance on Golden's (big) instances. In the other hand, some surveys covering exact algorithms, but often mainly devoted to heuristic methods, were presented by Christofides et al. [3], Bodin et al. [5], Laporte [22], Fisher [6], Toth and Vigo [8] and Golden et al. [9].

### III. PROBLEM FORMULATION

The formulation of the problem of routing vehicles can be defined as follows. Let  $G = (V, A)$  a complete graph with  $V = n + 1$  vertices, divided into two groups  $V = V^{DEP} \cup V^{CST}$ . The unique vertex  $v_0 \in V^{DEP}$  represents the depot where the products must be distributed and a fleet of  $m$  identical vehicles with limited capacity  $Q$  departs and returns.

The vertices  $v_i \in V^{CST}$  represent the customer indexed with  $i$ , where  $i = \{1, \dots, n\}$ , which require a service that is characterized by a non-negative demand associated  $q_i$ . The arcs  $a_{ij} \in A$  represent all possible connections from a node  $v_i$  to another node  $v_j$  with a service cost equal to  $c_{ij}$ . The main objective of the problem is to define a set of routes that allow every set of connections to dispatch the customer demand without exceeding the maximum capacity of the vehicles so that each sum of the costs associated to the established connections represent the minimum path.

The general VRP consist of the construction of a group of maximum  $m$  routes that satisfy the following requirements: (i) each route  $m$  must start and end at the depot; (ii) all customer demands must be met; (iii) the capacity of the vehicle must not exceed the capacity limit  $Q$ ; (iv) each client should be visited only once by a single vehicle or route.

As a mathematical representation of the problem, a binary linear programming model is proposed for the capacitated vehicle routing problem that is presented as an optimization problem. Its objective function is associated to the minimization of the cost or the total distance traveled by all vehicles, while restrictions associated to the visit of each customer by a single vehicle per route and the vehicle's capacity constraints are met. The model presented below corresponds to a modified version of the one proposed by Toth and Vigo in [8]. In this case variables can assume the value of 1 if the arc  $(i, j)$  is part of the solution and 0 otherwise.

The way the model is proposed is widely used for both the problem of VRP with symmetrical distances and VRP with asymmetric distances. The formulation has the following notation:

Sets

$V$	Set of customers indexed by $i, j$ and an unique depot indexed by $\{0\}$
$K$	Set of vehicles indexed by $k$
$S$	Set of proper subset of all nodes $V$

Constants

$c_{ij}$	The distance from node $i$ to node $j$
$q_i$	The demand at node $i \in V \setminus \{0\}$
$Q$	The loading capacity of the vehicle

Decision Variables

$x_{ij}$	Binary decision variable, taking a value of 1 if the vehicle goes from node $i$ to node $j$ and 0 otherwise
$u_i$	The remaining load when the vehicle leaves node $i$

$$\sum_{i \in V} \sum_{j \in V} c_{ij} x_{ij} \tag{1}$$

$$\sum_{j \in V, j \neq i} x_{ij} = 1 \quad \forall i \in V \tag{2}$$

$$\sum_{j \in V, j \neq i} x_{ij} - \sum_{j \in V, j \neq i} x_{ji} = 0 \quad \forall i \in V \tag{3}$$

$$\sum_{j \in V} x_{0j} = 1 \quad (4)$$

$$\sum_{j \in V} x_{0j} = |K| \quad (5)$$

$$u_j \leq u_i - q_j x_{ij} + Q(1 - x_{ij}) \quad \forall i, j \in V \setminus \{0\}, i \neq j \quad (6)$$

$$u_0 = Q \quad (7)$$

$$u_i \geq 0 \quad \forall i \in V \setminus \{0\} \quad (8)$$

$$x_{ij} \in \{0, 1\} \quad \forall i, j \in V \quad (9)$$

The objective function (1) minimizes the total cost. This value varies depending on the distances involved in each particular problem in the connections of the customers and the depot. Equations (2) and (3) ensure that each customer is visited only once. Constraints (3) allows the continuous flow between all nodes of the problem, which is both functional connections between clients and the routes with the depot. While (4) and (5) allow the depot has output connections with customers of the problem and that each of these connections does not exceed the maximum number of vehicles  $k$  of the problem, respectively. Constraints (6) record a vehicle's remaining load level based on node sequence. If the vehicle visits node  $i$  right after node  $j$ , the first term in the right-hand-side reduces the vehicle remaining load after leaving node  $i$  based on the demand at node  $i$  (if node  $i$  is the candidate site node, the demand at node  $i$  is zero). Otherwise, constraints (6) are relaxed. Constraints (7) ensure that the remaining load of vehicle is equal to its capacity  $Q$  when it leaves the depot. The remaining load must be nonnegative in constraints (8). And the condition (9) defines the connection variables  $x_{ij}$  as binary, because it can only assume the value of 0 or 1.

On the other hand, the problem of traveling salesman (TSP), which is addressed as a main sub-problem to test the methodology of implicit enumeration 0-1 or Balas's algorithm, has a lower level of complexity compared to the classical VRP, because their characteristics are similar, but they differ in their main objective, which in the case of the TSP, is focused on the connection of every customer through a single minimum path. The formulation of the TSP is presented as follows:

$$\sum_{i \in V} \sum_{j \in V} c_{ij} x_{ij} \quad (10)$$

$$\sum_{j \in V, j \neq i} x_{ij} = 1 \quad \forall i \in V \quad (11)$$

$$\sum_{j \in V, j \neq i} x_{ij} - \sum_{j \in V, j \neq i} x_{ji} = 0 \quad \forall i \in V \quad (12)$$

$$\sum_{i \in S} \sum_{j \in S} x_{ij} \leq |S| - 1 \quad (13)$$

$$x_{ij} \in \{0, 1\} \quad \forall i, j \in V \quad (14)$$

Since the sub-tours constraints (13) represent a huge amount of constraints for the implicit enumeration 0-1 method, the use of a subroutine that identifies the sub-tours

within each route is proposed. If necessary, the heuristic creates and adds the specific restrictions to remove the sub-tours, so it is not necessary to consider the numerous sub-tour constraints associated to each route.

#### IV. PROPOSED METHODOLOGY

In this section, a modified genetic algorithm of Chu-Beasley [23], [24], [25] is proposed (algorithm 1), and combined with a set of heuristics and exact techniques which would enable the algorithm to find solutions to the capacitated vehicle routing problem (CVRP) using the variant that works with homogeneous fleet and an unlimited number of vehicles. For this, the genetic algorithm must start with an initial population, which is built by using different kinds of heuristics that can generate individuals with good objective function (following (1)) and high level of diversity. For the selection process, two tournaments are held between a variable number of randomly selected individuals in the population. This gives the winners a chance to continue to the crossover stage. The resulting offspring from the crossover process goes through a stage for mutation step through local searches. As a stage of improvement, the exact method of implicit enumeration 0-1 is used to repair the intra-route codification, which is achieved by solving the Travel Salesman Problem (TSP) that is part of a CVRP sub-problem. Finally, the individual is presented to the population through the diversity and quality criteria. The process is carried out iteratively until reaching the stop criterion that was defined by a maximum number of iterations.

##### A. Initial population: heuristic process

For the generation of the initial population of the algorithm, is necessary to use a constructive methodology. For this specific case, the constructive proposed is represented by three heuristics that solve the problem in different ways: the generation of a big tour, the multiple solutions of the CVRP starting from each customer and multiple solutions with a factor of randomness. Hence, multiple characteristics of good quality individuals may remain in the execution of the algorithm.

The set of three heuristics are executed iteratively to generate individuals to fill the portion of the initial population that corresponds to each one of them or until the maximum number of individuals that can generate the heuristic is reached. Based on this, a set of heuristics capable to solve the CVRP was developed, for this case, the nearest neighbor algorithm (NN) for the low computational cost and a modified savings algorithm (MSA) as the one proposed by Subramanian [26] and in other proposals as Shang and Bouffanais[27][28]. The MSA was modified by adding randomness management, to provide a high level of diversity to the initial population. The third heuristic corresponds to the Lin-Kernighan-Helsgaun algorithm (LKH) [29] for the generation of a minimum path configurations with good quality.

For the stage of filling the population of the genetic algorithm, the proposed strategy its based on the generation of individuals depending on the specifics characteristics of each one of the heuristics used. In the case of LKH heuristic, the logic of a big tour is applied, where the TSP is solved with all the customers and without the depot using the Lin-Kernighan-

Helsgaun algorithm. The next step is the division of the big tour with the vehicle's capacity, starting from each customer, so the resulting set of routes for each customer as a starting point represents an individual of the population.

For the nearest neighbor algorithm, the complete CVRP is solved with all the nodes of the problem and starting every iteration from each customer, taking into account its demand and the maximum capacity of the vehicle. To build the routes of each individual, it is essential to make the respective connections only with the closest customer to the customer that is being analysed. In order to find feasible solutions, a variety of conditions must be met: (i) the next connection must be made with a node that has not been visited before; (ii) connections must respect the maximum capacity of the vehicle; (iii) The algorithm continues its search of routes, whenever there are customers who have not been visited. Otherwise, a new individual is generated with the routes found, a new client is selected to be the first customer visited and the route search starts again.

On the other hand, the modified savings algorithm is focused on generating individuals through partial solutions to the CVRP. However, in order to use all advantages of this algorithm, a two-stage strategy is proposed. The first stage is defined as the generation of the first individual, which is obtained by solving the CVRP with the original distances of the problem and following the methodology proposed by Subramanian [26], where it is possible to find an individual of good quality for the population. The next step was developed to fill the portion of the population that is assigned to the heuristic or in some cases, fill the remaining portion of the initial population (small instances). To make this possible, the algorithm must work with modified distances, which are constantly varied to add randomness to individuals. For this, the distances  $c_{ij}$  between all nodes of the problem is varied by a random factor  $\alpha = \{0.00, 0.05, 0.10, \dots, 1.65, 1.70\}$  that varies in each iteration. The modified distance  $c'_{ij} = \alpha * c_{ij}$ , and wherein each iteration is generated a set of modified distances to determine routes that represent partial solutions that will become part of the initial population.

### B. Selection

For the selection of individuals from the population, two stages are responsible to define the two best individuals, the result of two tournaments which are carried out with the random selection of a variable number of parents who enter to be compared by their respective objective function [23]. The winner of the tournament will continue with its respective population's identifier to be taken to the crossover stage.

### C. Crossover

Based on the fact that the individuals resulting from the selection stage may be considered as a permutation of the problem's customers, it is necessary to analyse different recombinant techniques, in order to adapt our crossover in the best way possible to take full advantage of the routes that have each parent. For that reason, recombinant techniques were reviewed, such as, the crossover with as a single point, two-point crossover and multi-point [23] crossover. They were also analysed as crossover, where specialized routes or parts

of routes are exchanged between individuals. One of the main objectives of this step is to allow propagation of the attributes of good-quality parents present in the crossover step. To make this possible, it is necessary to follow a sequence of steps that are based on the random selection of one of the two parents (algorithm 2), and then continue with the logic presented below:

- 1) *Select a random route of the selected parent.*
- 2) *Pass a complete route (if possible) or any part of the parent's route that are not included in the offspring's sequence.*
- 3) *Select the other parent and repeat until all the offspring is complete.*

### D. Mutation

The offspring obtained from the crossover sequence is subjected to a variable number of mutations, corresponding to typical movements for local search studied in literature, focused on search solution space through intra-route [30] changes. For this step, two stages allow the algorithm 3 do a search in the neighborhoods of solution space and perform a number of permutations of the offspring's configuration. In this way, it is possible to add diversity to the population and avoid being trapped in local optima prematurely. In this solution, instead of the typical movements reported by literature, a set of specialized movements is employed. Every move is performed only if the change leads to an improvement in the objective function. The mentioned movements are listed below:

- **Shift (1, 0):** A customer is transferred from its original route to another location in a different route.
- **Shift (2, 0):** Two consecutive clients are transferred from their current positions inside the route to other positions in different routes, keeping the intermediate arc that connect them.
- **Swap (1, 1):** Two different customers of different routes exchange their position.
- **Swap (2, 1):** The movement of an unbalanced exchange, obtaining a change of a pair of consecutive customers with one customer of a different route, where the intermediate arc in the pair of consecutive customers is preserved.
- **Swap (2, 2):** This movement follows the logic applied in the Swap (1, 1) resulting in the exchange of customers between consecutive pairs and two different routes, where the arc that connects each pair of customers is preserved.
- **Cross:** The exchange between two route segments.
- **Multi-Shift (1, 0):** A perturbation process is performed, inserting (insertions such as, Shift (1, 0)) random customers from their original routes to different routes, consecutively.
- **Multi-Swap (1, 1):** Two random customers are exchanged from their original routes a random number of times.

**E. Improvement Stage: Methodological enumeration 0-1 or implicit algorithm Balas**

The main reason for using specialized 0-1 implicit enumeration algorithms, which usually combined the technique of nested Benders decomposition to solve problems with integer and continuous variables, is that, unlike what happens with linear programming (LP), there is no integer programming algorithm that allows to solve all types of problems, because the efficiency of a LP algorithm depends largely on the particular characteristics of the problem. Therefore, there is no LP algorithm that is better than the other to solve all problems, or computer programs available with such features.

In this context, one of the first stages of a research project is to select a method to solve the problems resulting LP, and this selection should be made considering the different methods that exist in the literature such as: cutting-plane method, Branch & Bound, implicit enumeration, suboptimal methods, etc. At a later stage, a specialized algorithm should be developed that exploits the specific characteristics of the problem (algorithm 4).

*Notation:*

- +j indicates that the variable  $x_j = 1$
- j indicates that the variable  $x_j = 0$

An underscore as  $\underline{j}$ ,  $x_j$  element indicates that the variable to the alternative  $x_j = 0$  and was explored and probed.

**Partial solution (J):** It is an ordered set which values are defined as binary to a subset  $J \subseteq N$ . For example:

$$J = \{6, -2, \underline{-4}, 5\}$$

**Free variables (N - J):** They are variables that have not yet assigned a binary value to a partial solution and, therefore, is available to assume a value 0-1.

**Complement of J:** The set of solutions obtained from J assigning all variables that are still free binary values 0-1.

**Partial pruned solution:** A partial solution and can be pruned if all complement of J can be discarded for not being interesting.

**Glover's implicit enumeration scheme:** *In the algorithm of Balas, the best possible solution is stored. The list of  $2^n$  possible solutions is analyzed implicitly or explicitly, the last best feasible solution found, named incumbent, is the optimal solution.*

**1) Pruning tests**

The pruning tests are designed to exclude the maximum possible complements (i.e., derived solutions) of a partial solution for neither improving the objective function nor the feasibility (not interesting solutions). These tests are mainly of heuristic type therefore they may be so weak that allow explicit enumeration of almost all  $2^n$  feasible solutions or so powerful that exclude virtually all solutions not interesting.

At iteration t,  $J_t$  is the partial solution, then we have:

$$S_i^t = b_i - \sum_{j \in J_t; j > 0} a_{ij}; i \in M \quad (15)$$

$$z^t = \sum_{j \in J_t; j > 0} c_j \quad (16)$$

where  $S_i^t$  defines the value of the slack variables and  $z^t$  the objective function. Be  $z_{min}$  the best feasible solution found, so-called incumbent. The idea of probing tests, in an attempt to exclude a set of possible solutions because they are not considered interesting, can be based on two basic considerations. For a partial solution defined by  $J_t$  assume that  $S_i^t < 0$  for at least one  $i \in M$ .

**Test 1 (Balas):**

It defines:

$$A_t = \{j \in (N - J_t) \mid a_{ij} \geq 0, \forall i \mid S_i^t < 0\} \quad (17)$$

The elements in  $A_t$  are those free variables that being raised to 1 do not improve the infeasibility of the current partial solution.

Be:

$$N_t^1 = (N - J_t) - A_t$$

If  $N_t^1 = \emptyset$  it means that no free variables can be raised to 1, then  $J_t$  is pruned by infeasibility and a backward movement must be performed.

**Test 2 (Balas):**

It is a test for exclusion of variables, in the vector of free variables, using an optimization criterion, where:

$$B_t = \{j \in N_t^1 \mid (z^t + c_j) \geq z_{min}\} \quad (18)$$

The elements  $B_t$  free variables are those that improve the infeasibility of the problem. Each variable carries an objective function greater than  $z_{min}$  incumbent and this is a worst quality objective function. Thus,  $x_j, j \in B_t$  are excluded as candidates to assume a value of 1 because they are not interesting for optimality.

Be:

$$N_t^2 = N_t^1 - B_t$$

If  $N_t^2 = \emptyset$ ,  $J_t$  is pruned because it has no better complement feasible and must performed a backward movement.

**Test 3 (Balas):**

It is a pruning test, where:

$$C_t = \left\{ i \in M \mid S_i^t < 0; \sum_{j \in N_t^2} a_{ij}^- > S_i^t \right\} \quad (19)$$

$$a_{ij}^- = \min(0, a_{ij})$$

If  $C_t \neq \emptyset$ , at least one restriction will remain infeasible,

then  $J_t$  is pruned, because it has no feasible complement. It must be performed a backward movement. If  $C_t = \emptyset$ , the pruning test continue.

**Test 3' (Glover-Zionts):**

This test assesses whether each variable  $x_j \in N_t^2$  that is promoted to 1 will cause an increase in the objective function



beyond the relationship permissible for the restriction violated [31].

For each  $S_i^t < 0$  calculate:

$$r_i = \min_{j \in N_t^2} \left\{ \left( \frac{S_i^t}{a_{ij}} \right) C_j; a_{ij} < 0 \right\} \quad (20)$$

Where  $\left( \frac{S_i^t}{a_{ij}} \right) C_j$  is the relative cost of each variable  $x_j$ , regarding the degree of reducing the infeasibility of that variable in the violated constraint  $i$ . If  $r_i \geq (z_{min} - z^t)$  then  $J_t$  is pruned.

### 2) Forward movements

When pruning tests fail, the number of variables in  $J_t$  should be increased. That is, some free variables must take defined values and should be included in the set  $J_t$  assuming specific values either 0 or 1.

#### Test of Geoffrion:

This expansion test attempts to set values of the free variables into 0 or 1 in order to ensure feasibility, identifying essential variables for the problem [32]. However, this does not guarantee that the movement promote variables to the partial solution (may fail). The test is formulated as follows:

For each  $i$  such that  $S_i^t < 0$  and every  $j \in N_t^2$ , if:

$$\left[ S_i^t - \sum_{j \in N_t^2} \min(0, a_{ij}) - |a_{ij}| \right] < 0 \quad (21)$$

Then:

$$x_j = 0 \text{ if } a_{ij} > 0 \text{ or } x_j = 1 \text{ if } a_{ij} < 0$$

#### Test of Glover-Zionts:

This test allows a free variable set at 0 through

a very simple test [33], [31] formulated as follows:

For each  $i$  and  $p \in N_t^2$  such that  $a_{ip} > S_i^t$ , calculate,

$$c_h = \min\{c_j \mid j \in N_t^2 - \{p\}, a_{ij} < 0\} \quad (22)$$

If  $c_h + c_p \geq z_{min} - z^t$  then  $x_p = 0$

#### Test of Balas:

It selects a variable  $x_{j^*}$ ,  $x_{j^*} \in N_t^2$  to take the value of 1 and add  $j^*$  to  $J_t$ . This variable is selected by the relationship:

$$v_{j^*} = \max_{j \in N_t^2} \{v_j\} \quad (23)$$

Where:

$$v_j = \sum_{i \in M} \min(0, S_i^t - a_{ij}); j \in N_t^2 \quad (24)$$

If  $v_{j^*} = 0$ , there is no infeasibility and so  $J_{t+1}$  is feasible and this new solution should lead to better value  $z_{min}$ , the incumbent. So  $J_{t+1}$  is pruned and the incumbent must be updated.

### 3) Backward move

If a partial solution has been pruned, then a backward move must be made in the scheme of Balas. This means that

the value of one of the variables of  $J_t$  must be modified. In the normal scheme of Balas, this process is performed with the LIFO (last-in, first-out) rule. With LIFO, the last variable to enter the list is the first considered for further exploration. For example, let  $J_t$ .

$$J_t = \{2, \underline{-5}, 3, 6, 1, \underline{4}\}$$

If  $J_t$  is pruned, the new partial solution is defined for the relation:

$$J_t = \{2, \underline{-5}, 3, 6, \underline{-1}\}$$

Evidently, the order in which variables are analysed alters the process enumeration. This fact was noted by Tuan who showed it is not necessary to strictly follow the LIFO [34] rule implicitly embedded in the numbering scheme Glover, to ensure proper development of numbering scheme Glover. Suggests Tuan determining a subset  $J''$  consists of variables that can be selected for the development of future scans, this that is, those variables that can be underlined. If  $j_1$  is the element  $J_t$  would be selected by the rule LIFO, then the elements of  $J''$  are those elements that are located  $J_t$  from  $j_1$  even to find the first highlighted in a section of element right to left in the elements of  $J_t$ .

### F. Methodology for sub-tours identification

The methodology of implicit enumeration has a set of pruning and expansion movements, which allow it to realize a reduction of the search space of the problem. However, to solve the TSP is necessary to use a model full of the problem with all sub-tour constraints that prevent isolated customer groups in the final solution.

This group of constraints also represents a huge burden on the constraint matrix of the method. Therefore, the determination of a subroutine responsible for identifying isolated customer groups was necessary.

As a result, it was found that the implicit enumeration algorithm ensured a global optimum if all sub-tour constraints are included for small and medium instances. Nevertheless, at a high computational cost. Hence, a modified and relaxed model without sub-tours constraints was implemented instead of using the extensive one. To make that possible, a subroutine to identify sub-tours that would reduce the number of sub-tour restrictions that were strictly necessary was added. This achieves a good compromise between processing time and quality responses.

## V. PSEUDOCODE OF THE PROPOSED METHODOLOGY

### Algorithm 1. GACBIE

---

```

procedure GACBIE(Instance)
    LKH ← LinKernighanTSP(Customers)
    NN ← NearestNeighborTSP(Customers)
    MS ← ModifiedSavingsTSP(Customers)
    Population ← GeneratePopulation(LKH, NN, MS)
    P* ← worstCromosome(Population)
    F* ← f(P*)
    *P ← bestCromosome(Population)
    *f ← f(*P)
    P ← ∅
    f ← ∞
    for i ← 0 to MaxIter do

```

---

---

```
P ← Crossover(Population, seed)
P ← Mutation(P, seed)
SubTour ← ∅
while isFeasible(P) ≠ TRUE do
    Model ← GenerateReducedModel(P, SubTour)
    P ← BalasAlgorithm(Model, P)
    SubTour ← findSubTour(P)
end while
if diversityCriteria(P) = TRUE then
    if (f(P) < F*)
        AddToPopulation(Population, P, P*)
        P* ← worstCromosome(Population)
        F* ← f(P*)
    end if
else
    if (f(P) < *f)
        AddToPopulation(Population, P, *P)
        *P ← P
        *f ← f(*P)
    end if
end if
end for
bestP ← findBest(Population)
return bestP
end procedure GACBIE
```

---

#### Algorithm 2. Crossover

---

```
procedure Crossover(Population, seed)
    ParentSet_1 ← SelectRandomParents(Population, seed)
    ParentSet_2 ← SelectRandomParents(Population, seed)
    Paren_1 ← FindBestObjectiveFunction(ParentSet_1)
    Paren_2 ← FindBestObjectiveFunction(ParentSet_2)
    while AllCustomersIncluded(P) ≠ TRUE do
        R ← SelectRandomRoute(Parent_1, Parent_2)
        P ← P ∪ R - (P ∩ R)
    end while
    return P
end procedure Crossover
```

---

#### Algorithm 3. Mutation

---

```
procedure Mutation(s, seed)
    s' ← s
    iter ← 0
    s* ← s
    f* ← f(s)
    while iter to MaxMutIter do
        s ← LocalSearch(s)
        if f(s) < f(s') then
            s' ← s
            iter ← 0
        end if
        iter ← iter + 1
        s ← Perturbation(s, seed)
    end while
    if f(s') < f* then
        s* ← s'
        f* ← f(s')
    end if
    return s*
end procedure Mutation
```

---

#### Algorithm 4. BalasAlgorithm

---

```
procedure BalasAlgorithm (Model, Path)
    M ← Model
    Jt ← InitialSolution(M, Path)
    NodeStack ← ∅
    NodeStack ← Jt
    while NodeStack ≠ ∅ do
        if isFeasible(Jt, M) then
            if f(Jt) < f* then
                f* ← f(Jt)
                J* ← Jt
                Jt ← backwardMove(Jt, M, NodeStack)
            else
                if PruningTest(Jt) = TRUE then
                    Jt ← backwardMove(Jt, M, NodeStack)
                else
                    Jt ← forwardMove(Jt, M, NodeStack)
                end if
            end if
        end if
    end while
    return J*
end procedure BalasAlgorithm
```

---

## VI. COMPUTATIONAL EXPERIMENTS

As a final step, computational tests are done to analyse the performance, behavior and contribution of the GACBIE methodology. Results are compared to BKS (Best Known Solution).

The implementation of the proposed methodology was done in C++ language under the G++ compiler that is part of a set of free license compilers from the collection of GNU compilers (GCC). The operating system used was Ubuntu 14.04 with kernel version 3.14, on an Intel Core i5 3.2 GHz and 8 GB RAM machine. Below there is a comparison between the best solutions obtained in the literature, for instances of CVRP, as also implementation of the various proposals for implicit enumeration 0-1 method.

### A. Parameter setting

The correct behavior of the proposed algorithm is directly dependent on the parameters used, which is the reason to make a parameter assignment, so that in this way it is possible to determine a range in which the methodology can vary with a semi-random form.

In the development process of the algorithm, some parameters come to dominate the behavior of the algorithm. This is because some stages as mutation and improvement processes are highly structured, while the selection step depends on the number of individuals into competition. In this regard, some authors like [36] warn that a large number of individuals can make the algorithm too elitist. So the values assigned are intended to allow the range to preserve the diversity of algorithm individuals avoiding taking very good quality repetitively.

In addition, portions of the initial population assigned to the heuristics are presented in an equitable way for the LKH and NN, because they are algorithms that makes low use of computing resources.

While the modified savings algorithm, has a much smaller portion of the population since the algorithm can take much longer to generate individuals, given their division in stages, there is a very high probability that the best individual who can be generate by the algorithm, is present in the generation of the first stage.

TABLE I. PARAMETER SETTING

Description	Range/Value
Population size	60
LKH	40%
NN	40%
MSA	20%
Selection	2-3
MaxIter	100-1000

Table I shows, respectively, the values for: the population size, the corresponding portions of the Lin-population Kernigan heuristics (LKH), nearest neighbor (NN) and the modified savings algorithm (MSA), the selection range (the number of individuals entering the tournament) and MaxIter

range (iterations number range), which may be increased depending on the size of each instance.

B. Computational results

1) GACBIE vs GACB and GACBC

Table II summarizes the results for three algorithms: the GACBIE (with implicit enumeration method as step improvement), the genetic algorithm with CPLEX® as improvement step (GACBC) and the same genetic algorithm without implicit enumeration but heuristics (nearest neighbor and simple inter-routes exchange) as improvement stage (GACB). All the algorithms were tested with 20 instances proposed by Augerat *et al.* (1995) (available at [35]) and compared with BKS to measure performance and effectiveness in the search for the optimal response. In the first two columns of Table 2, the instance name used followed by the number of customers (n) followed by the results of tests with the GACBIE, GACBC and GACB, accompanied by the GAP (denoted by equation (25)) between the BKS and the best solution reached each methodology, the average responses and their execution time. Finally The last column represents the number of iterations for execution.

TABLE II. RESULTS OF THE TEST ON INSTANCES OF AUGERAT ET AL

Instance	n	BKS	GACBIE				GACBC				GACB				Iterations	
			Best	Avg.	GAP(%)	Time (min)	Best	Avg.	GAP(%)	Time (min)	Best	Avg.	GAP(%)	Time (min)		
A-n32-k5	32	784	784	784.00	0.0	6.12	784	784.00	0.0	1.12	784	784	0.0	0.11	300	
A-n33-k5	33	661	661	662.00	0.0	6.40	661	662.00	0.0	1.40	662	663.89	0.15	0.09	300	
A-n33-k6	33	742	742	742.00	0.0	5.58	742	742.00	0.0	2.58	742	742.78	0.0	0.08	300	
A-n34-k5	34	778	778	778.08	0.0	5.81	778	778.00	0.0	1.81	778	779.05	0.0	0.08	300	
A-n36-k5	36	799	799	810.15	0.0	6.41	799	816.06	0.0	2.41	817	822.47	2.25	0.13	300	
A-n37-k5	37	669	669	674.07	0.0	8.66	669	674.71	0.0	3.66	677	681.26	1.196	0.15	300	
A-n37-k6	37	949	949	956.26	0.0	9.71	949	952.00	0.0	6.71	960	963.02	1.159	0.16	600	
A-n38-k5	38	730	730	745.56	0.0	10.32	730	747.00	0.0	8.32	756	759.67	3.562	0.17	600	
A-n39-k5	39	822	822	833.60	0.0	18.84	822	834.00	0.0	9.84	842	843.11	2.43	0.24	600	
A-n39-k6	39	831	831	834.39	0.0	18.36	831	835.97	0.0	10.36	850	852.01	2.28	0.22	600	
A-n44-k6	44	937	946	955.09	1.07	20.26	946	953.34	1.07	12.26	952	954.86	1.60	0.34	600	
P-n16-k8	16	450	450	450.00	0.0	1.13	450	450.00	0.0	0.12	450	452.42	0.0	0.02	200	
P-n19-k2	19	212	212	212.00	0.0	1.30	212	212.00	0.0	0.10	212	212	0.0	0.04	200	
P-n20-k2	20	216	216	216.00	0.0	2.25	216	216.00	0.0	0.13	216	220.01	0.0	0.04	200	
P-n21-k2	21	211	211	211.00	0.0	2.05	211	211.00	0.0	0.20	211	215.62	0.0	0.05	200	
P-n22-k2	22	216	216	218.35	0.0	2.58	216	218.35	0.0	0.60	216	225.36	0.0	0.05	200	
P-n22-k8	22	590	590	593.84	0.0	2.23	590	593.14	0.0	0.23	590	595.12	0.0	0.05	200	
P-n23-k8	23	529	529	529	0.0	2.80	529	529	0.0	0.40	529	529.87	0.0	0.06	200	
P-n40-k5	40	458	462	462	0.8	18.18	462	462	0.8	10.11	462	504.9	0.04	0.30	600	
P-n45-k5	45	510	523	525.65	2.5	26.8	523	520.06	2.5	9.06	523	613.5	0.15	0.37	600	
			<b>Avg. GAP</b>	<b>0.2185</b>				<b>Avg. GAP</b>	<b>0.2185</b>				<b>Avg. GAP</b>	<b>0.74085</b>		

$$gap = \frac{Best_{Sol} - BKS}{BKS} * 100 \tag{25}$$

2) NIC vs NIRSD

The strategy to reduce sub-tours constraints for routes of VRP that are repaired by the methodology of implicit enumeration, proves to be an interesting idea to reduce computational time for large problems. As observed in Table III, the computational time reduction that occurs with the increase of size of the problems is due to the implicit enumeration methodology is responsible for testing each of the possible connections possible that cannot be pruned for their selection strategies variables.

To revise the advantage of using a strategy of identifying sub-tours, tests were conducted where it was possible to make a comparison of computation times between the method of implicit enumeration with the full model including restrictions on sub-tours (NIC), the same method of implicit enumeration constrained dynamic sub-tours (NIRSD) and the speed up gap (denoted by equation (26)), wherein the steps of perturbation and crossover alter the continuity of each of the routes of the customer involved.

$$gap = \frac{NIC\_time}{NIRSD\_time} \tag{26}$$

TABLE III. NIC VS NIRSD RESULTS

Test	n	NIC(s)	NIRSD(s)	GAP(times)
1	5	2.02	5.5	-0.36
2	6	5.76	8.12	-0.7
3	7	8.04	10.28	0.78
4	8	18.22	14.97	1.21
5	9	37.19	28.43	1.31

## VII. CONCLUSIONS

At the end of the proposed project, a series of results found through the tests conducted in the modified genetic algorithm method relied on a technique of implicit enumeration stage 1-0 improvement implemented for the routing problem of trained vehicles with homogeneous fleet with diverse neighborhoods and a set of perturbation techniques. Results could be obtained the following observations:

- Replacing the typical crossover combinations by a crossover based on route exchange proved to be a very efficient strategy to keep good characteristics of the routes obtained in the population.
- By using the implicit enumeration technique the use of commercial solvers is avoid, which is an advantage in terms of cost for the implementation.
- Compared to those techniques that have to solve matrixes (Simplex, for instance) the Balas's technique is less complex since it has to realize simple algebraic operations.
- Computational experiments show that the algorithm proposed is able to obtain high quality solutions within reasonable computing times.
- Adding all restrictions on sub-tours is prohibitive for the slave problem as shown in the table III. It is more convenient to add only sub-tours constraints necessary.
- The results suggest that the proposed algorithm can be applied to other variants of problem routing vehicle, considering the inclusion of multiple depots, and use fleets heterogeneous vehicles, and the possible addition of techniques they can further narrow is the solution space.
- The use of genetic algorithm with CPLEX® compared with the methodology of implicit enumeration, shows equivalent results in reasonable computer times, considering that the solver CPLEX® has strong pruning techniques, allowing the solver to reduce the dimension of the solution space in a quickly and efficiently way. Compared with the proposed methodology, it shows a very good performance since the technique achieves global optimal in traveling salesman problems assigned. Furthermore, the use of the genetic algorithm with simple heuristics movements as improvement stage demonstrate the need for a strong methodology to repair the connections that can be generated at the stage of mutation and crossover.

## REFERENCES

- [1] G. B. Dantzig and J. H. Ramser, "The truck dispatching problem," *Management Science*, vol. 6, pp. 80–91, 10 1959. 195
- [2] C. E. Miller, A. W. Tucker, and R. A. Zemlin, "Integer programming formulation of traveling salesman problems," *J. ACM*, vol. 7, no. 4, pp. 326–329, Oct. 1960. [Online]. Available: <http://doi.acm.org/10.1145/321043.321046> 195
- [3] Christofides, N., M;ingozzi A. Toth P. (1981) "Exact algorithms for the Vehicle Routing Problem,based on Spanning Trees and Shortest Path Relaxations". *Mathematical Programming* 20.Pp 255-282.
- [4] Christofides, N., M;ingozzi A. Toth P. (1981) "Exact algorithms for the Vehicle Routing Problem,based on Spanning Trees and Shortest Path Relaxations". *Mathematical Programming* 20.Pp 255-282.
- [5] Bodin L., Golden B, Assad and M. Ball, (1983) "Routing and scheduling of vehicles and crews: The state of the art", *Computers and OperationsResearch*, 10(2),, 63-211.
- [6] Fisher, M. and Jaikumar, R. (1981).A generalized assignment heuristic for vehicle-routing. *Networks*, 11(2):109–124.
- [7] G. Laporte, H. Mercure, and Y. Nobert, "An exact algorithm for the asymmetrical capacitated vehicle routing problem," *Networks*, vol. 16, no. 1, pp. 33–46, 1986. [Online]. Available: <http://dx.doi.org/10.1002/net.3230160104> 195
- [8] P. Toth and D. Vigo, Eds., *The Vehicle Routing Problem*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2001. 195, 197
- [9] Golden L., T. Wong. (1981) "Capacitated arc routing problems." *Networks*, 11:305-315.
- [10] R. Romero and A. Monticelli, "A zero-one implicit enumeration method for optimizing investments in transmission expansion planning," *Power Systems*, *IEEE Transactions on*, vol. 9, no. 3, pp. 1385–1391, Aug 1994. 196, 199
- [11] Clarke, G., Wright, J., "Scheduling of vehicles from a central depot to a number of delivery points". *Operations research*, 1964 12 (4), 568–581.[online] available: <http://pubsonline.informs.org/doi/abs/10.1287/opre.12.4.568>
- [12] A. Pessoa, M. P. de Aragão, and E. Uchoa, "Robust branch-cutand-price algorithms for vehicle routing problems," in *The Vehicle Routing Problem: Latest Advances and New Challenges*, ser. *Operations Research/Computer Science Interfaces*, B. Golden, S. Raghavan, and E. Wasil, Eds. Springer US, 2008, vol. 43, pp. 297–325. [Online]. Available: [http://dx.doi.org/10.1007/978-0-387-77778-8\\_14](http://dx.doi.org/10.1007/978-0-387-77778-8_14) 195
- [13] G. Laporte, Y. Nobert, S. Taillefer, "A branch-and-bound algorithm for the asymmetrical distance-constrained vehicle routing problem, *Mathematical Modelling*", Volume 9, Issue 12, 1987, Pages 857-868, [Online]. Available: [http://dx.doi.org/10.1016/0270-0255\(87\)90004-2](http://dx.doi.org/10.1016/0270-0255(87)90004-2).
- [14] Christofides, N., and S. Eilon. "An Algorithm for the Vehicle-dispatching Problem". *OR* 20.3 1969: 309–318. [online] available: <http://www.jstor.org/stable/3008733>
- [15] Christofides, N., Mingozi, A., Toth, P., 1979. The vehicle routing problem. In: Christofides, N., Mingozi, A., Toth, P., Sandi, C. (Eds.), *Combinatorial Optimization*. Vol. 1. Wiley Interscience, pp. 315–338.
- [16] Rochat, Y., Taillard, E. D., "Probabilistic diversification and intensification in local search for vehicle routing". *Journal of heuristics* 1, 1995 (1), 147–167. [online] available: <http://dx.doi.org/10.1007/BF02430370>
- [17] Fukasawa, R., Longo, H., Lysgaard, J., Poggi de Aragao, M., Reis, M., Uchoa, E., Werneck, R.. Robust branch-and-cut-and-price for the capacitated vehicle routing problem, *Mathematical programming*, 2006, vol. 106, 3, pp. 491–511.
- [18] Baldacci, R., Christofides, N., Mingozi, A., 2008. An exact algorithm for the vehicle routing problem based on the set partitioning formulation with additional cuts. *Mathematical Programming* 115 (2), 351–385
- [19] Contardo, C., Martinelli, R., 2014. A new exact algorithm for the multi-depot vehicle routing problem under capacity and route length constraints. *Discrete Optimization* 12, 129–146.

- [20] Røpke, S., 2012. Branching decisions in branch-and-cut-and-price algorithms for vehicle routing problems. Presentation in Column Generation 2012.
- [21] Vidal, T., Crainic, T. G., Gendreau, M., Prins, C., 2014. A unified solution framework for multi-attribute vehicle routing problems. *European Journal of Operational Research* 234 (3), 658 – 673.
- [22] Laporte, G. (2007) What you should know about the vehicle routing problem. *NavalResearch Logistics* 54, 8, 811-819
- [23] P. C. Chu and J. E. Beasley, “A genetic algorithm for the generalized assignment problem,” *Comput. Oper. Res.*, vol. 24, no. 1, pp. 17–23, Jan. 1997. [Online]. Available: [http://dx.doi.org/10.1016/S0305-548\(96\)00032-9](http://dx.doi.org/10.1016/S0305-548(96)00032-9) 199, 201
- [24] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Cambridge, MA, USA: MIT Press, 1992. 199
- [25] G. Syswerda, “Schedule optimization using genetic algorithms,” in *Handbook of Genetic Algorithms*, L. Davis, Ed. New York, NY: Van Nostrand Reinhold, 1991. 199
- [26] A. Subramanian, P. H. V. Penna, E. Uchoa, and L. S. Ochi, “A hybrid algorithm for the heterogeneous fleet vehicle routing problem,” *European Journal of Operational Research*, vol. 221, no. 2, pp. 285 – 295, 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0377221712002093> 200
- [27] Y. Shang and R. Bouffanais, “Consensus reaching in swarms ruled by a hybrid metric-topological distance,” *Eur. Phys. J. B*, vol. 87, no. 12, pp. 1–8, 2014.
- [28] Y. Shang and R. Bouffanais, “Influence of the number of topologically interacting neighbors on swarm dynamics,” *Sci. Rep.*, vol. 4, p. 4184, 2014.
- [29] S. Lin and B. W. Kernighan, “An effective heuristic algorithm for the traveling-salesman problem,” *Operations Research*, vol. 21, no. 2, pp. 498–516, 1973. [Online]. Available: <http://dx.doi.org/10.1287/opre.21.2.498> 200
- [30] P. Penna, A. Subramanian, and L. Ochi, “An iterated local search heuristic for the heterogeneous fleet vehicle routing problem,” *Journal of Heuristics*, vol. 19, no. 2, pp. 201–232, 2013. [Online]. Available: <http://dx.doi.org/10.1007/s10732-011-9186-y> 202
- [31] F. Glover and S. Zionts, “A note on the additive algorithm of balas,” *Operations Research*, vol. 13, no. 4, pp. 546–549, 1965. [Online]. Available: <http://dx.doi.org/10.1287/opre.13.4.546> 206, 207
- [32] A. M. Geoffrion, “Integer programming by implicit enumeration and balas’ method,” *SIAM Review*, vol. 9, no. 2, pp. 178–190, 1967. [Online]. Available: <http://www.jstor.org/stable/2027441> 206
- [33] F. Glover, “A multiphase-dual algorithm for the zero-one integer programming problem,” *Operations Research*, vol. 13, no. 6, pp. 879–919, 1965. [Online]. Available: <http://dx.doi.org/10.1287/opre.13.6.879> 207
- [34] N. P. Tuan, “A flexible tree-search method for integer programming problems,” *Operations Research*, vol. 19, no. 1, pp. pp. 115–119, 1971. [Online]. Available: <http://www.jstor.org/stable/168870> 208
- [35] “NEO Research Group,” <http://neo.lcc.uma.es/vrp/vrp-instances/capacitated-vrp-instances/>, accessed: 2014-01-12. 211, 213
- [36] Rendon, R.A.G. and Zuluaga, A.H.E. and Ocampo, Eds., *Técnicas metaheurísticas de optimización*, Universidad Tecnológica de Pereira, 2008. 100, 150, 9789587220070.

# Automation of Optimized Gabor Filter Parameter Selection for Road Cracks Detection

Haris Ahmad Khan

Laboratoire Electronique,  
Informatique et Image (Le2i)  
Université de Bourgogne  
Dijon, France

M. Salman, Sajid Hussain

Wah Engineering College,  
University of Wah  
Wah Cantt, Pakistan

Khurram Khurshid

Electrical Engineering Department  
Institute of Space Technology  
1, Islamabad highway, Islamabad,  
44000, Pakistan

**Abstract**—Automated systems for road crack detection are extremely important in road maintenance for vehicle safety and traveler's comfort. Emerging cracks in roads need to be detected and accordingly repaired as early as possible to avoid further damage thus reducing rehabilitation cost. In this paper, a robust method for Gabor filter parameters optimization for automatic road crack detection is discussed. Gabor filter has been used in previous literature for similar applications. However, there is a need for automatic selection of optimized Gabor filter parameters due to variation in texture of roads and cracks. The problem of change of background, which in fact is road texture, is addressed through a learning process by using synthetic road crack generation for Gabor filter parameter tuning. Tuned parameters are then tested on real cracks and a thorough quantitative analysis is performed for performance evaluation.

**Keywords**—Pavement Cracks; Automated detection; Gabor Filters; Genetic Algorithm; Parameter Selection

## I. INTRODUCTION

Transportation structure is built around pavements and roads. These roads require periodic health inspections so that information concerning their physical condition can be gathered and recommendations can be made for their reconstruction or repair if required. In time detection of cracks, which is pivotal in executing economical pavement maintenance thus avoiding more damage, both physical and financial. The topmost layer of the flexible pavement is known as asphalt course which is constructed by black bitumen and one down size aggregate. The mixing ratio, aggregate color is not always constant. That is why texture of pavement varies continuously after some distance.

Texture is an innately non-local image property. Texture is found in most natural images and therefore, the analysis of texture is an important topic of research in computer vision in the past decades [1]. Generally speaking, there are a number of filter configurations designed for texture segmentation. Most of filters are useful for particular texture but don't produce desired results when texture is varied.

Gabor filter is found helpful to analyze textures having definite orientation and frequency characteristics. Majority of prior research on Gabor filters is centered on filter-bank technique that need a filter set with predefined parameters for efficiently covering the frequency range. Such approach considerably affects classification results and is computationally expensive, so such technique is not

appropriate for pavement crack detection where texture is varying all the time. In this work we develop a system for designing optimized Gabor filter for detection of pavement cracks out of pavement texture.

Additionally, optimization of Gabor filter parameters is done in semi-supervised manner while crack detection is unsupervised, because natures of cracks that can come across to this system are not known. Thus, the objective is differentiating between background textures and defective feature (crack) automatically. As the pavement cracks are randomly shaped and can be oriented in any direction, there is no specific ground truth for the comparison of effectiveness of the system. We propose simulated ground truth creation of pavement crack for optimization of Gabor filter in varying texture. Synthetic ground truth is useful in many ways like, ease in controlling parameters, fast and convenient testing of datasets.

The rest of this paper is organized as follows. Section 2 is the description and review of related work in simulated ground truth generation. In Section 3, we present the theoretical background of Gabor filter, genetic algorithm and Structural Similarity Index Measure (SSIM). We present the model of our system for optimization of Gabor filter parameters in Section 4. In Section 5, we discuss experimental results. Concluding remarks are in Section 6.

## II. LITERATURE REVIEW

The classic standard for validation of the segmentation method is evaluation with manually done segmentation but, manual segmentation experiences the lack of reproduction and reliability, and different systems adopt different methods for manually marking part of interest. The true ground truth may require estimation from a set of manual segmentations. The condition for manual ground truth generation in such methods, is need of getting the data with all possible variations to guarantee the consistency of the training and test sets

Using simulated information has various advantages including speedy generation of dataset with low cost, management of degradations parameters and models, and convenient tests on equivalent data having same information but different background.

Simulated ground truth is formed by Nataliya et al for evaluation of performance in [2] for dirt particle counting and categorization. The aim is to create an image with a consistent



conditions and dirt particles of various types, so that the nature and the position of every particle are known.

In [3] Serranho et al introduced a scheme to create ground truth for computation of performance metrics in perspective of OCT image processing. The study focused to set up a noise-free simulated image that imitates actual OCT data. They propose that mean square error is computed with noise-free ground truth. The computation of simulated image depends on properties of actual OCT data.

Optical Character Recognition (OCR) systems require training on the characters which are to be recognized in test data for which ground truth data is needed. In [4] Beusekom et al used synthetic ground truth for calibration of system. This data is set of character images along with their ASCII codes. Among the approaches intended for generation of ground truth of actual data, one practice is scanning documents and then utilizing them. By means of a grouping technique, the features extracted from the test document are compared with test image. For making the scheme robust to transformations, creation of simulated ground truth for training is used as a technique.

In [5], David Doermann et al discuss the generation of ground truth for OCR using the method of generating simulated document images and representative ground truth files by means of metafile information and custom print driver. In [6] Sheraz et al proposed automatic word ground truth creation of camera captured document images using a document image retrieval system. In [7] Vivek et al created and tested a method for curve detection by means of synthetically formed curves. Validation and evaluation of magnetic resonance images (MRI) segmentation is a difficult task due to shortage of reliable ground truth information. In [8] Marcel et al, suggest a method for generation of simulated multi-modal 3D brain MRI with edema and tumor. Using such ground truth images having variations in tumor dimension, position, and degree of adjacent edema, segmentation techniques can be used on images for identification. The simulated ground truth MRI is used for validation entire brain's segmentation, together with gray matter, white matter, edema and tumor. Santuari et al [9], presents a real time graphical simulator based on client-server structural design. The rendering engine, supported by a specific client application for the automatic production of objective oriented motion of synthetic characters, is used to create realistic image sequences for wide-ranging performance evaluation of computer vision algorithms for person tracking.

### III. PARAMETER OPTIMIZATION

In this section, we discuss briefly about theoretical background of Gabor filter, and discuss method for optimization of parameters using genetic algorithm (GA) with Structural Similarity Index Measure (SSIM) as cost function.

#### A. Gabor Filter

Fourier analysis is the most commonly used tools in signal processing to analyze frequency characteristics of a particular signal. There are eight degrees of freedom in the 2D Gabor filter family:  $x_0$  and  $y_0$  specify the position of the filter in 2D spatial domain;  $u_0$  and  $v_0$  are modulation coordinates specify the position of the filter in 2D frequency domain,  $\theta$  is desired orientation,  $\omega$  is spatial frequency,  $\psi$  is the phase offset of the

modulation factor, which decides the symmetry or anti-symmetry of the filter and the width (a) and the length (b) of the elliptical Gaussian (2D) envelope and the angle between orientation of sinusoidal wave vector and the two dimensional Gaussian axes [22-23]. The general form of the Gabor filter is as in (1).

$$g(x, y) = s(x, y)w(x, y) \quad (1)$$

Where the Gaussian envelopes can be termed as:

$$s(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{1}{2}\left[\left(\frac{x'}{\sigma_x}\right)^2 + \left(\frac{y'}{\sigma_y}\right)^2\right]\right] \quad (2)$$

Here  $\sigma_x$  and  $\sigma_y$  are the scale factor of the neighborhood. The complex sinusoidal with phase offset  $\psi$  is:

$$w(x, y) = e^{j(2\pi\omega_0x' + \psi)} = \cos(2\pi\omega_0x' + \psi) + \sin(2\pi\omega_0x' + \psi) \quad (3)$$

So the overall 2D Gabor filter

$$g(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left[\left(\frac{x'}{\sigma_x}\right)^2 + \left(\frac{y'}{\sigma_y}\right)^2\right]} \times e^{j(2\pi\omega_0x' + \psi)} \quad (4)$$

Where  $x' = x\cos\theta + y\sin\theta$  and  $y' = -x\sin\theta + y\cos\theta$ . The real part of Gabor filter is as under

$$g_r(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left[\left(\frac{x'}{\sigma_x}\right)^2 + \left(\frac{y'}{\sigma_y}\right)^2\right]} \times \{\cos(2\pi\omega_0(x\cos\theta + y\sin\theta)) + \psi\} \quad (5)$$

And the imaginary component of the complex Gabor filter is

$$g_i(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left[\left(\frac{x'}{\sigma_x}\right)^2 + \left(\frac{y'}{\sigma_y}\right)^2\right]} \times \{\sin(2\pi\omega_0(x\cos\theta + y\sin\theta)) + \psi\} \quad (6)$$

Equation (5) and (6) can be written in terms of aspect ratio  $\gamma = \frac{\sigma_x}{\sigma_y}$  (i.e., ellipticity of the Gabor) and wavelength  $\lambda$  of the sinusoidal factor.

$$g(x, y) = \frac{\gamma}{2\pi\sigma^2} e^{-\frac{1}{2}\left[\frac{x'^2 + \gamma^2 y'^2}{\sigma^2}\right]} \times e^{j\left(\frac{2\pi x'}{\lambda} + \psi\right)} \quad (7)$$

#### B. Gabor Filter Power Spectrum

The outputs of a real and an imaginary Gabor filter for each image point can be combined in a single quantity which is termed as the Gabor energy. This feature is associated with the complex cell, which is the model of orientation selective neuron in the primary visual cortex [22] and the Gabor energy is closely related to the local power spectrum and is defined in the following way;

$$P(x, y) = E^2(x, y) = g_r^2(x, y) + jg_i^2(x, y) \quad (8)$$

A Gabor filter of wavelength of 20, variance = 4, phase offset = 0 is shown in Fig. 2. And the frequency response of the Gabor filter is illustrated in Fig. 2.

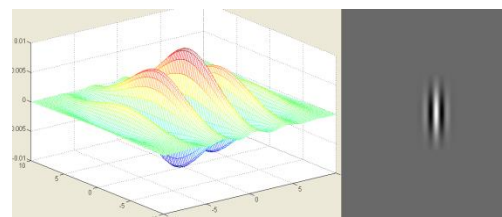


Fig. 1. Gabor filter at 0 degree orientation

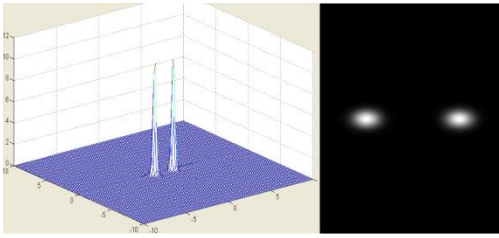


Fig. 2. Frequency response of filter in Fig. 1

### C. Genetic Algorithm

Genetic Algorithm (GA) begins by defining the optimization variables, cost function, and associated cost. Its ending is like other optimization algorithms as well, by evaluating for convergence. However, this algorithm is relatively dissimilar. Components of GA along with their paths are shown in figure 3. A cost function produces an output from set of input variables called a chromosome. This cost function can be a mathematical function [24]. The objective is to transform the output in a desirable manner by identifying the appropriate values for input variables. Cost function in proposed method is SSIM and input variables to be optimized for a given texture in pavement texture are variance  $\sigma$  and wavelength  $\lambda$ . In GA literature, the term fitness is broadly used to allocate the output of the objective function. Most of the optimization literature deals with minimization of cost while fitness implies a maximization problem.

The GA begins by defining a chromosome in form of an array of variables to be optimized which are known as Genes of a chromosome.

$$\text{Chromosome} = [\text{Gene1}, \text{Gene2}] = [\lambda, \sigma] \quad (9)$$

In this research, Gene1 is wavelength  $\lambda$  and Gene2 is variance  $\sigma$ , jointly are the parameters of the Gabor filter to be optimized and the desired output from GA gives the most suitable Gabor Filter to be used for crack detection in a particular texture of pavement.

Furthermore, most of the optimization problems require constraints or variable bounds while unconstrained variables can take any value. These bounds are defined in proposed algorithm section.

### D. Structural Similarity Index Measure (SSIM)

The structural similarity Index Measure (SSIM) [25] [26] [27] is a technique for computing the similarity between two images. SSIM is formulated as a development in quality assessment. Traditional techniques such as peak signal-to-noise ratio (PSNR) and mean square error (MSE) are proven to be inconsistent with human eye perception. The advantage of SSIM over other techniques mentioned earlier (MSE and PSNR) is that SSIM consider image degradation as perceived variation in structural information while other methods estimate perceived errors. Structural information is the scheme which suggests that the pixels have strong inter-dependencies particularly when they are close in spatial domain. These dependencies contain significant information concerning the structure of the objects in an image.

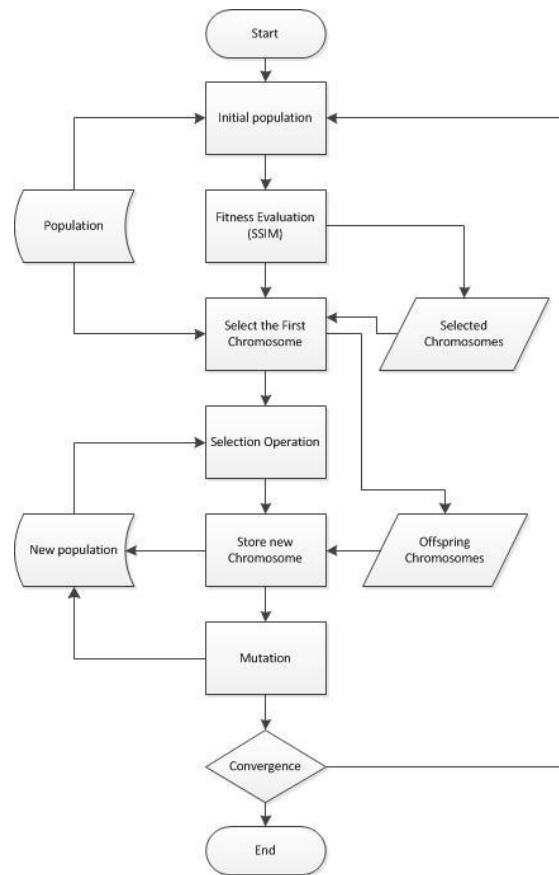


Fig. 3. Structure of Genetic Algorithm

The SSIM metric of an image is calculated on a variety of windows. To evaluate between two windows  $x$  and  $y$  of size  $N \times N$  is:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (10)$$

Where

$\mu_x$  is mean and  $\sigma_x^2$  is variance of  $x$ .

$\mu_y$  is mean and  $\sigma_y^2$  is variance of  $y$ .

$\sigma_{xy}$  is the covariance between  $x$  and  $y$ ;

$C_1$  and  $C_2$  are two variables for stabilizing division with weak denominator;

SSIM is applied on two images and the resultant SSIM index returns a decimal value between -1 and 1. Value of 1 is returned if both images are exactly the same. While -1 shows complete dissimilarity.

## IV. PROPOSED ALGORITHM

In this section, the proposed method for optimization of the parameters of the Gabor filter through simulated ground truth with the aid of Genetic algorithm is discussed. The proposed system is implemented on a personal computer using MATLAB.

### A. Simulated Ground Truth

To increase the robustness of the optimization of Gabor parameters, the technique of simulated ground truth is used. This ground truth is simulated from within the texture of test image after selecting a 100 by 100 patch containing flexible pavement texture but no crack. There are relatively few low intensity pixels as compare to mean of whole patch. Low intensity pixels contain the crack present in image that is why it is avoided. This 100 by 100 patch is fused with another patch of same size which contains a synthetically created multi oriented crack as shown in figure 4.



Fig. 4. Synthetic crack

The synthetic crack is arbitrary, thin as compare to original cracks and most importantly is stretched almost every direction. All these characteristics make this crack appropriate for fusion in real texture of road. As mentioned earlier that intensity values of cracks are low as compare to the rest of pavement texture but the texture is not uniform and changes continuously, the intensity must be less than the average intensity of the image. The flow diagram for simulated ground truth is shown in figure 5.

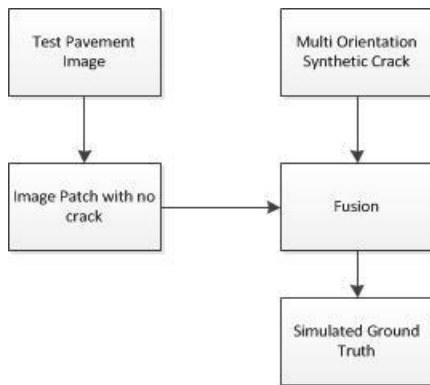


Fig. 5. The process of creating the ground truth

The input test image shown below has a multi directional crack and high coarseness. The 100 by 100 patch of pavement with no crack and the fused crack on the patch is shown in figure 6.

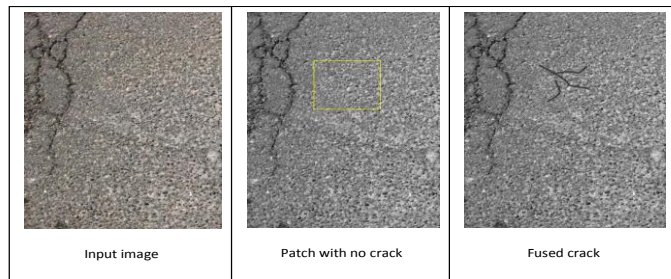


Fig. 6. Original and synthetic cracks

The 100 by 100 sized simulated ground truth is shown below in figure 7.



Fig. 7. Simulated Ground Truth

### B. Initial Population

The designed chromosome for this research contains two genes. The first gene is the wavelength of the Gabor filter while the second gene is the variance of the Gabor filter. These two variables are the key parameters to define a Gabor filter. To get the optimized values of these variables for any type of pavement texture with lesser computation time the variable are constrained to some initial and final values. The variance of Gabor filter must be greater than 2 and lesser than 8 and the wavelength must be from 8 to 30. These constrain are selected after intense experiments. To define these genes the binary Genetic algorithm is used. The figure 8 below shows one of the chromosomes in the initial population.

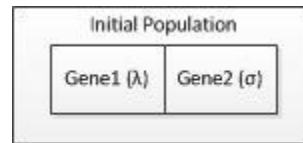


Fig. 8. Chromosomes with genes

### C. Fitness Evaluation

As described earlier that SSIM is one of the best measure of the image structure similarities. To compute the cost of each chromosome in a generation, the Gabor filter created by each chromosome is applied to the ground truth. The resultant image is then compared with the crack image as shown in fig [2]. Higher the similarity between the image the greater will be the SSIM values. It must be noted that the SSIM values varies from 1 to -1.

### D. Convergence

As soon as the stopping criteria are fulfilled, the optimized values of Gabor filter are extracted from the fittest chromosome. Now these values are fed into the crack detection algorithm to extract the crack from the original test image. As described earlier the texture of pavement varies after some distance, but it varies very little in neighboring patches. So the optimized parameters can be used for some distance and will be updated after some distance. This will increase the efficiency of system by decreasing the computation time. The figure 9 illustrates the complete scenario of the proposed method.

### E. Crack Detection Algorithm

An optimized filter bank is generated using the optimized Gabor filter parameters. The filter bank contains multi orientation filter. Generally, the number of orientations depends on the application where the Gabor filter is applied.



As the number of orientations is increased, the output is more accurate [27]. But increase in orientations also increases the computational time and complexity of the system. The Gabor filter of a given orientation is then convolved with the input preprocessed image. After the completion of convolution, the real part of the response out of the filter is threshold to produce a binary image. Finally, the binary images resulting from the differently oriented filters are combined by an OR operation to produce an output image that contains detected crack segments. The whole process is shown in flow diagram (figure 9b). In Gabor filter bank the  $\lambda$  and  $\sigma$  are the optimized values from GA.

V. EXPERIMENTATION AND RESULTS

The input image is first selected from image dataset. By using this image a simulated ground truth is created as shown previously in figure 7. Now the ground truth and the crack image are then fed into the Genetic algorithm with the cost function based on the SSIM. Initial population size is selected randomly as 20, selection rate 0.5 and the mutation rate is 0.01. The optimized parameters after 25 generation are variance 2.08 and the wavelength 12.35 which are very close to the manually obtained parameters. Now the crack detection algorithm is applied to extract the crack. After experimentation, the image is thresholded at 0.7 out of 1. The output image Fig. 10 shows the extracted crack. The Small square tiles of  $10 \times 10$  pixel size are consider for crack revelation and the result are based only on that tiny region. Another extracted pavement crack is shown in figure 13.

The output of the system is computed by the precision and recall of the manually selected and optimized Gabor filter not by the values of the variance and wavelength of the optimized Gabor filter. The graph below in figure 12 shows the precision of five pavement crack images

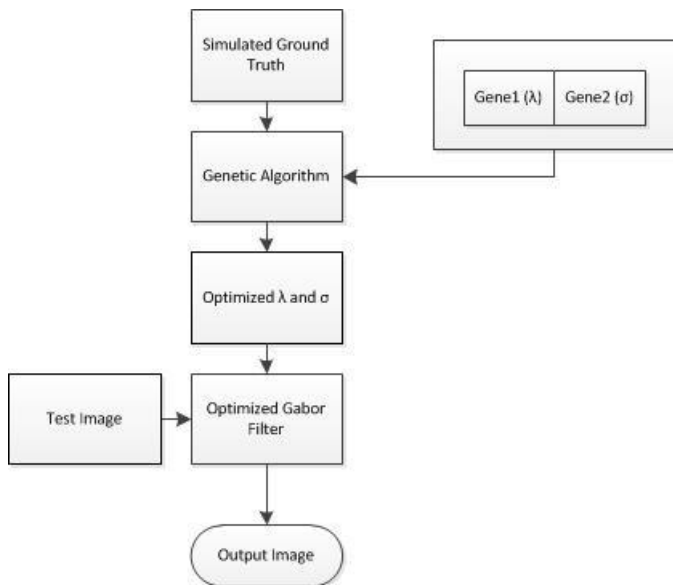


Fig. 9. a. Proposed Method Flow Chart

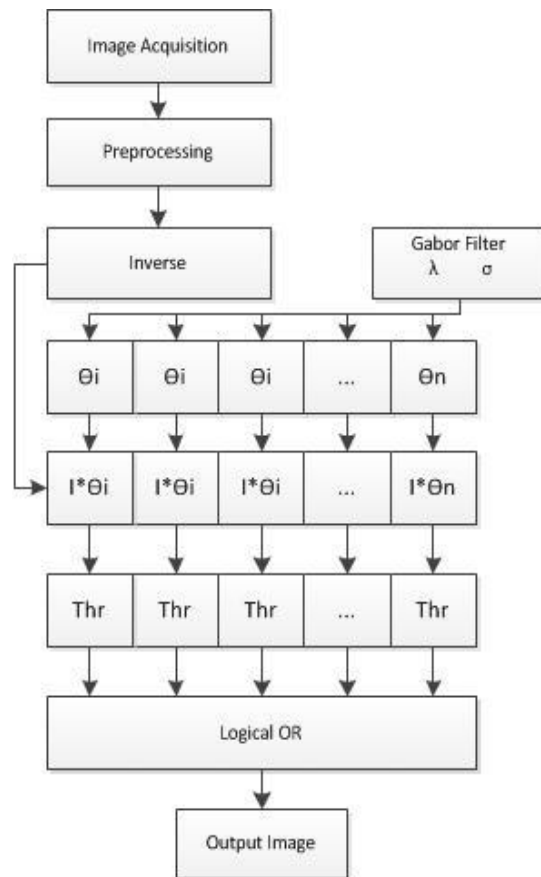


Fig. 8. b. Crack Detection algorithm (Flow chart)

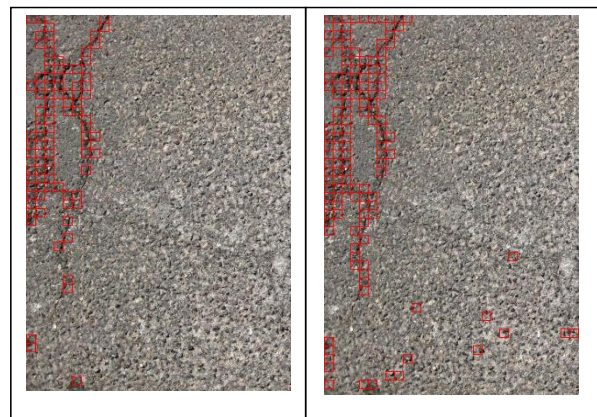


Fig. 9. Left: manually selected Gabor filter extracted crack, Right: crack extracted by optimized Gabor filter

Figure 12 shows the comparison between the output manual and optimized Gabor filter.

Table 1 shows the results for 5 images. The first two are industrial images acquired by the LRIS system. The last three are captured by the Canon IXUS 80 IS camera. The total number of crack tiles and the non-crack tiles are based on manual estimation.

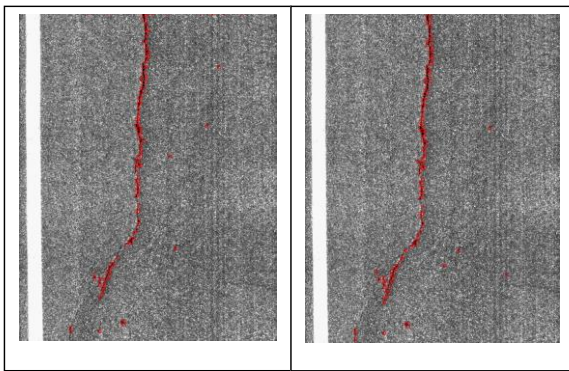


Fig. 10. Left: manually selected Gabor filter extracted crack, Right: crack extracted by optimized Gabor filter

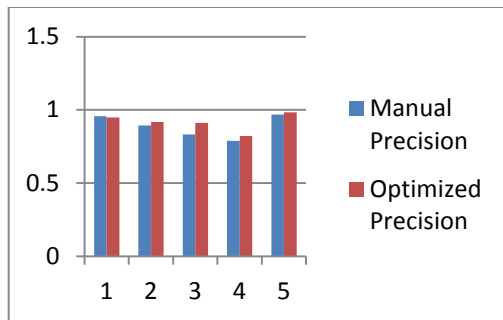


Fig. 11. Comparison between manual and Optimized Precision

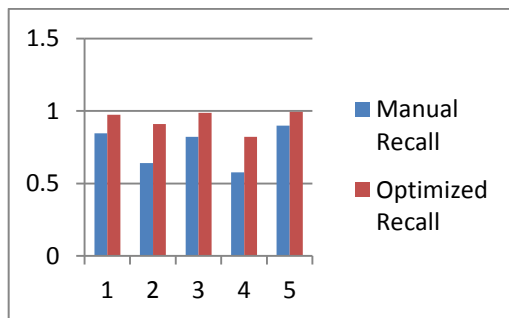


Fig. 12. Comparison between manual and Optimized Precision

## VI. CONCLUSION

This paper proposes a method to automatically optimize Gabor filter for road crack detection. The experimental results show that though the precision of the manually selected and Automated optimized Gabor filter may become similar, the execution time of optimization through our proposed method is very less and thus can be used for practical applications. As texture of pavement is changed after some interval, parameters of the Gabor filter are also updated and optimized so that pavement crack detection process can be performed smoothly and without manual calibration.

### REFERENCES

- [1] Lazebnik, S.; Schmid, C.; Ponce, J., "A Sparse Texture Representation Using Local Affine Regions," Pattern Analysis And Machine Intelligence, IEEE Transactions on , Vol.27, No.8, Pp.1265,1278, Aug. 2005.
- [2] N. Strokina, A. Mankki, T. Eerola, L. Lensu, J. Käyhkö, And H. Kälviäinen, "Semisynthetic Ground Truth for Dirt Particle Counting and Classification Methods", Proc. MVA, 2011, pp.215-218.
- [3] Serranho, P. Maduro, C. Santos, T. Cunha-Vaz, J. Bernardes R., "Synthetic OCT Data for Image Processing Performance Testing," 18th IEEE International Conference on Image Processing (ICIP), pp 401-404, Sept. 2011.
- [4] Beusekom, Joost van; Shafait, F.; Breuel, T.M., "Automated OCR Ground Truth Generation," Document Analysis Systems, 2008. DAS '08. The Eighth IAPR International Workshop on , vol., no., pp.111,117, 16-19 Sept. 2008
- [5] Gang Zi; Doermann, D., "Document image ground truth generation from electronic text," Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on , vol.2, no., pp.663,666 Vol.2, 23-26 Aug. 2004
- [6] Ahmed, S.; Kise, K.; Iwamura, M.; Liwicki, M.; Dengel, A., "Automatic Ground Truth Generation of Camera Captured Documents Using Document Image Retrieval," Document Analysis and Recognition (ICDAR), 2013 12th International Conference on , vol., no., pp.528,532, 25-28 Aug. 2013
- [7] Kaul, Vivek; Yezzi, Anthony; Tsai, Yichang, "Detecting Curves with Unknown Endpoints and Arbitrary Topology Using Minimal Paths," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.34, no.10, pp.1952,1965, Oct. 2012
- [8] M. Prastawa, E. Bullitt, and G. Gerig, "Synthetic Ground Truth for Validation of Brain Tumor MRI Segmentation", Proc. MICCAI, 2005, pp.26-33.
- [9] Santuari and O. Lanz and R. Brunelli, "Synthetic movies for computer vision applications," Proc. 3rd IASTED International Conference: Visualization, Imaging, and Image Processing (VIIP 2003) pp. 1-6
- [10] R. W. Lucky, "Automatic equalization for digital communication," Bell Syst. Tech. J., vol. 44, no. 4, pp. 547-588, Apr. 1965.
- [11] De Valois, R. L., Albrecht, D. G. and Thorell, L. G., "Spatial-frequency selectivity of cells in macaque visual cortex", Vision Research 22, 545-559 (1982)
- [12] S. P. Bingulac, "On the compatibility of adaptive controllers ." in Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory, New York, 1994, pp. 8-16.
- [13] G. R. Faulhaber, "Design of service systems with priority reservation," in Conf. Rec. 1995 IEEE Int. Conf. Communications, pp. 3-8.
- [14] Doyle, W.D., "Magnetization reversal in films with biaxial anisotropy," Magnetics, IEEE Transactions on , vol.2, no.2, pp.68,73, Jun 1966
- [15] Juette, G.W.; Zaffanella, L.E., "Radio Noise Currents and Audible Noise on Short Sections of UHV Bundle Conductors," Power Apparatus and Systems, IEEE Transactions on , vol.PAS-89, no.5, pp.902,913, May 1970
- [16] J. G. Kreifeldt, "An analysis of surface-detected EMG as an amplitude-modulated noise," presented at the 1989 Int. Conf. Medicine and Biological Engineering, Chicago, IL.
- [17] J. Williams, "Narrow-band analyzer", Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.
- [18] N. Kawasaki, "Parametric study of thermal and chemical non-equilibrium nozzle flow," M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.
- [19] J. P. Wilkinson, "Nonlinear resonant circuit devices (Patent style)," U.S. Patent 3 624 12, July 16, 1990.
- [20] Jain, A. K., Ratha, N. K.; Lakshmanan, S., "Object detection using Gabor filters", Pattern Recognition 30, 295-309 (1997).
- [21] Porat, M. and Zeevi, Y. Y., "The generalized Gabor scheme of image representation in biological and machine vision", IEEE Trans. Pattern Analysis Mach. Intell. 10, 452-468 (1988)
- [22] Daugman, J. (1980)., "Two-dimensional analysis of cortical receptive field profiles", Vision Research, 20, 846-856
- [23] Daugman, J. (1985), "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters", Journal of the Optical Society of America-A, 2(7), 1160-1169.
- [24] Shi, Xian-Jun; Hong Lei, "A Genetic Algorithm-Based Approach for Classification Rule Discovery," Information Management, Innovation Management and Industrial Engineering, 2008. ICIMI '08. International Conference on , vol.1, no., pp.175,178, 19-21 Dec. 2008

- [25] Zhou Wang; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P., "Image quality assessment: from error visibility to structural similarity," Image Processing, IEEE Transactions on, vol.13, no.4, pp.600,612, April 2004.
- [26] Salman, M. Mathavan, S.; Kamal, K.; Rahman, M., "Pavement crack detection using the Gabor filter," Intelligent Transportation Systems - (ITSC), 2013 16th International IEEE Conference on , vol., no., pp.2039,2044, 6-9 Oct. 2013
- [27] Casti, P., Mencattini, A., Salmeri, M., & Rangayyan, R. M., "Analysis of Structural Similarity in Mammograms for Detection of Bilateral Asymmetry," Medical Imaging, IEEE Transactions on, 34(2), 662-671, 2015.

TABLE I. RESULT FOR 5 IMAGES

Img No.	Crack Tiles	True Positives		False Positives		False Negatives		True Negatives	
		Manual	Optimized	Manual	Optimized	Manual	Optimized	Manual	Optimized
1	155	131	148	6	8	24	4	11181	11179
2	145	93	132	11	12	52	13	1120	1119
3	73	60	72	12	7	13	1	1367	1372
4	45	26	37	7	8	19	8	432	431
5	173	151	169	5	3	17	1	11169	11169



# Classification of Hand Gestures Using Gabor Filter with Bayesian and Naïve Bayes Classifier

Tahira Ashfaq, Khurram Khurshid

Electrical Engineering Department, Institute of Space Technology  
1, Islamabad Highway, Islamabad, 44000, Pakistan

**Abstract**—A hand Gesture is basically the movement, position or posture of hand used extensively in our daily lives as part of non-verbal communication. A lot of research is being carried out to classify hand gestures in videos as well as images for various applications. The primary objective of this communication is to present an effective system that can classify various static hand gestures in complex background environment. The system is based on hand region localized using a combination of morphological operations. Gabor filter is applied to the extracted region of interest (ROI) for extraction of hand features that are then fed to Bayesian and Naïve Bayes classifiers. The results of the system are very encouraging with an average accuracy of over 90%.

**Keywords**—Human Computer Interaction; Hand Segmentation; Gesture recognition; Gabor Filter; Bayesian and Naïve Bayes classifiers; Feature Extraction; Image Processing

## I. INTRODUCTION

These days artificial intelligence is being commonly used in various daily life applications. Image processing and machine learning play a vital role in artificial intelligence applications. Systems like face recognition, facial expression recognition and gesture recognition involve multiple steps of image segmentation and processing and research in this domain continues to prosper to find out more suitable solutions for forthcoming challenges and problems.

Gesture recognition is considered as quite a difficult job due to nature of challenges that hinder the state of the art systems get 100% accurate results. Different techniques like Support Vector Machines (SVM), Neural Networks (NN) and Hidden Markov Models (HMM) are used by researchers for classification of gestures in various classes. Hand gesture classification has various applications, like in Human Computer Interactive applications, gesture recognition and classification allows users to interact efficiently with computers using easy to use/understand hand gestures. William1995 was one of the first ones that established the use of hand gesture recognition for controlling household machines. Ayesha2010 defines that in gesture recognition main goal is to create a system which can recognize precise human gestures for relaying useful information.

The hand gestures can be categorized as static hand gestures and dynamic hand gestures. Static hand gestures are

basically still images having a certain position and features of hand. The hands can be differentiated by different hand position features in which finger positions are determined and features are found for the precise finger, thumb and palm pattern. Thus a static hand gesture is signified by an individual frame or image.

Dynamic hand gestures on the other hand can be differentiated through the starting and last stroke movement of moving hand gesture. A video is the best example of dynamic hand gesture. Mainly two different approaches are used for hand gesture-tracking named as glove based systems and vision based systems for hand gesture analysis. In the first approach, sensors and other instruments are attached with user's hand in subjective analysis. In order to analyse data, Magnetic field tracker device and data glove or body suits are used. Extracted features from different human body positions such as the angle and rotation are essential for examination in such systems. Vision-based approach is the substitute to glove based technique because normally it does not need sophisticated or even small external equipment apart from a camera. These systems work using various image processing and machine learning algorithms. In any system however, the 3 important steps include hand region segmentation, feature extraction and then final classification & recognition.

## II. PROPOSED SYSTEM

In this paper, we present a scheme for static gesture recognition using vision based approach. Our proposed system can be summarized in the following steps. In the first step pre-processing is done, where initial pre-filtering for noise removal is performed followed by conversion of images from RGB to LAB color space. Then the third dimension grey level is converted to binary using global threshold and is mapped to extract just the desired hand region. For removal of undesirable areas, combination of different morphological operations is applied on the image. The resulted image is then used to extract the edges of the segmented region. Once the hand region is segmented, we can extract the certain features of hand which are then used in the classification module for gesture mapping. The classification is done by using Bayesian and naïve Byes classifier and their performance is compared in the end.

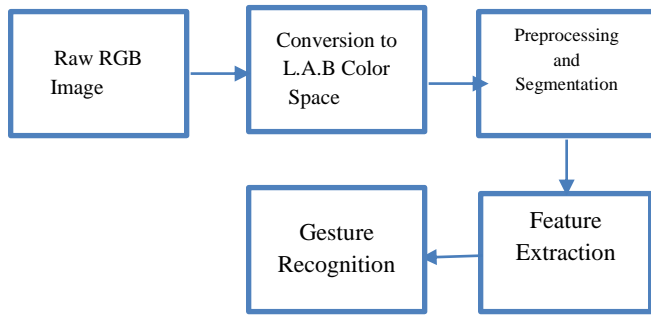


Fig. 1. System Block Diagram

The procedures included in above block diagram are discussed in detail in the following section.

### A. Preprocessing

#### 1) Conversion to L.A.B Space

We use RGB color images as input which is converted to LAB color space by analysing the pixel's color and differentiating on its basis. Skin color has certain properties that can be extracted easily via color information. In contrast to RGB and CMYK, L.A.B color is designed for accurate estimations and therefore better results in case of image segmentation.

Luminance  $L^*$  is separated by color space from two different color components ( $a^*$  and  $b^*$ ). This is because human perception of lightness is almost matched by the 'L' component. Accurate colored balance correction can be made by modifying output curves in 'a' and 'b' components, or for the adjustment of lightness and contrast using the 'L' component.

#### 2) Thresholding

Automatic grey thresholding of third component 'b' is performed in L.A.B. color space.

$$I(:, :, 3) \quad : \quad I \text{ represents the input image}$$

The obtained gray image is converted to binary using the below mentioned threshold. The resultant binary image can be mapped to original image to view the segmented region.

$$J(x, y) = \begin{cases} I(x, y), & \text{Where } O(x, y) \sim 0 \\ 0, & \text{otherwise} \end{cases}$$

$J$  in above equation represents the output of segmented image.  $O$  is the binary image and if it has a nonzero pixel value then the same spatial coordinates intensity of  $I$  will be assigned to the output pixel  $J(x, y)$ . Thus segmented skin area can be extracted through this mapping.

### B. Segmentation

For segmentation, we have to perform the following operations.

#### 1) Morphological Operations

Opening is performed in order to remove narrow joints and noise. In Opening we first perform erosion by using  $5 \times 5$  structuring element.

$$A \ominus B = \bigcap_{b \in B} A_{-b}$$

Here erosion is used to remove the extra region. Where  $A$  represents the image and  $B$  denotes the structuring element. In the above equation, structuring elements are applied on the image  $A$ . If the 'ON' pixels of a structuring element hit the pixel image where the pixel value is more likely to be 0, then the center value of mask is shifted towards the new image. Whenever there is a missing 'ON' pixel in the structure element, either structure element or mask will be shifted towards the next pixel.

After erosion we dilate image by using  $5 \times 5$  structuring element in order to fill the disturbed region.

$$A \oplus B = \bigcup_{b \in B} A_b$$

Here  $A$  is represents the gray scale image, and  $B$  again denotes the structuring element.

To recap, dilation adds pixels to the object boundary while erosion removes object pixels from boundary.

#### 2) Canny edge detector

After morphological operations, we need to detect the boundary of the region. For that we tried various known filters like Sobel, Prewitt, Canny. We finalized Canny edge detection as it gave more desirable results as compared to the other two options. After this stage, we get a well detected hand region and its boundary.

### C. Feature Extraction

Once the hand region is extracted, we have to extract good features that have more discriminative power. These precise features are used for classifying various hand gestures so that the problems like interclass similarity and intra class variability can be catered for easily. Extraction of the these best features is definitely a challenging task. Various researchers have used various features for classifying gestures based in internal information of the object in image. A good feature set highly effects the performance of a classifier.

For our technique, we use Gabor filter based features that tell us about the nature of gesture using its pattern and orientation. The advantage we have using Gabor filter is that we get large variance between the features of different classes which minimizes the chances of error if used properly in a classifier.

#### 1) Gabor Filter

Gabor filter has been used in literature for various applications including satellite image segmentation, urban zone detection [17], document and camera images for edge detection etc. In spatial domain, a 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave. The impulse response of Gabor filter is defined by a sinusoidal wave. Sinusoidal wave is known as a plane wave for 2D Gabor filters, multiplied by a Gaussian function. Because of the multiplication-convolution property (Convolution theorem), the Fourier transform of a Gabor filter's impulse response is the convolution of the Fourier

transform of the harmonic function and the Fourier transform of the Gaussian function. Gabor filter has a real and an imaginary component representing orthogonal directions. The two components may be formed into a complex number or used individually.

Complex equation is given below

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi \frac{x'}{\lambda} + \psi\right)\right)$$

Real part is as follows

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \psi\right)$$

And imaginary part is as follows

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \sin\left(2\pi \frac{x'}{\lambda} + \psi\right)$$

Where

$$x' = x \cos(\theta) + y \sin(\theta)$$

$$y' = -x \sin(\theta) + y \cos(\theta)$$

In the above equation

The wavelength of sinusoidal factor is represented by  $\lambda$ .

$\theta$  is the orientation of the normal to the parallel stripes of the Gabor function,  $\psi$  is the phase offset,  $\sigma$  is the sigma/standard deviation of Gaussian envelope, and  $\gamma$  is the spatial aspect ratio that specifies the Ellipticity of the support for Gabor function.

Parameter	Symbol	Values
Orientation	$\theta$	$\{0, \frac{\pi}{8}, \frac{1\pi}{8}, \frac{2\pi}{8}, \frac{3\pi}{8}, \frac{4\pi}{8}, \frac{5\pi}{8}, \frac{6\pi}{8}, \frac{7\pi}{8}\}$
Wavelength	$\lambda$	$\{4, \sqrt[4]{2}, 8, \sqrt[3]{2}, 16\}$
Phase	$\psi$	$\{0, \frac{\pi}{2}\}$
Gaussian envelope	$\gamma$	$\gamma = \lambda$

We apply this filter to each hand image to find out the subjected feature values from the image.

#### D. Classification

Classification means to segment regions of interest into various groups called classes. Mainly there are two types of classification, Supervised and Unsupervised.

In unsupervised classification (also known as clustering in pattern recognition) the segmented groups are without labels, while in supervised classification we need some training samples that are helpful in classifying the unknown data or test cases. The grouping or matching is based on certain features that are extracted from the objects. These features should be discriminating in nature to give a correct and meaningful classification of the data. In our case, we give the Gabor Filter based features to two classifiers (Bayesian and Naïve Bayes) for deciding the class of the query gesture.

#### 1) Bayesian classifier

Bayesian classifier is based on calculating posterior probabilities using Bayes theorem. Posterior probabilities which are also known as class density estimates are calculated as below:

$$P(H | X) = \frac{P(X | H)P(H)}{P(X)}$$

In the above expression, posterior probability is denoted by  $P(X | H)$  in which X is conditioned on H. The prior probability of X is denoted by  $P(X)$ . Classifier's accuracy highly depends on this parameter.

#### 2) Naïve Bayes Classifier

The other technique that we tried was the Naïve Bayes Classification. This classifier basically estimates the class of unknown data item using probability models. Using Bayes theorem, posterior probability can be calculated using the following expression

$$P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)}$$

Here  $P(C_i)$  is known as class prior probabilities and it can be estimated by the following expression.

$$P(C_i) = \frac{S_i}{S}$$

Where  $S_i$  represents the selected training sample, which has class  $C_i$  and  $S$  denotes the total number of training samples.

### III. EXPERIMENTAL RESULTS

Now for testing the above explained methodology, we have used data with complication or real background. The data is acquired by a 7 mega pixel camera. We have a total of ten hand gestures that we are classifying. In our data set, we have collected 18 images of each hand gesture in which 5 are used for training and 13 for final testing. To justify the best accuracy of our proposed result, we have compared our technique results with a similar work done by nawazish et.al (2013) [18] that used Neural Network Based hand gesture recognition of dumb people. In that paper, the authors worked on a technique using neural networks with the combination of mean and entropy feature. In the paper he extracted features using block base and used 4 blocks of the image and extracted the feature separately. After extracting the features he used an array to store the results of all blocks and for classification he used neural networks. The results are given below in table 1. The table shows that our method of Gabor Filter based feature extraction and Bayesian Classification works well and classifies the gestures 90% of the time while the method of Nawazish achieved a classification rate of 88.25%. Step wise results of various stages of our technique are shown in figure 2. Two examples are depicted here in detail.

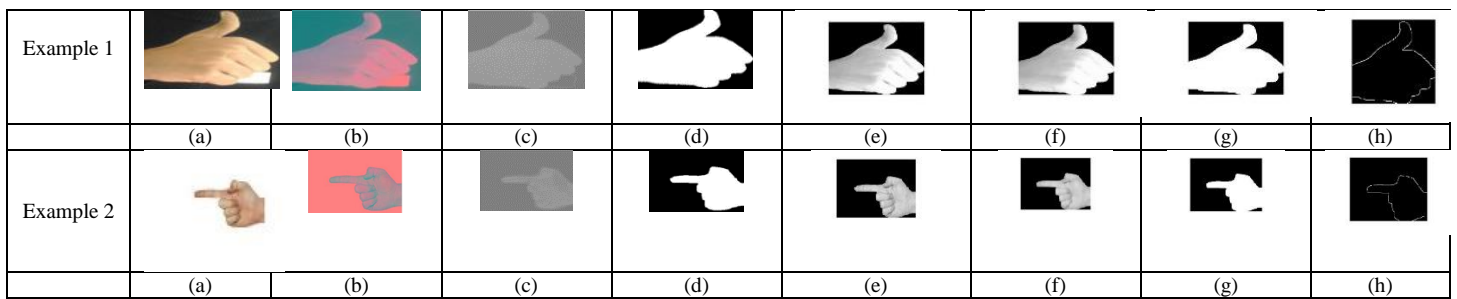


Fig. 2. Segmentation Results (a) Original image (b) LAB color space image (c) third color component of Lab image A (,.;, 3) (d) binary Image (e) Binary mapping (f) Morphological Operation opening (g) Final Binary image (h) Segmented Hand

TABLE I. COMPARISON OF RESULTS

Gestures	Comparison of various techniques		
	Nawazish et.al (2013)	Naïve Bayes	Bayesian Classifier
Stop	89.0	90.12	90.53
Pointing	85.2	87.20	88.01
Self-Pointing	87.34	89.01	89.85
Drinking	90.05	90.12	91.01
Take Care	89.66	90.00	91.00
.....	.....	.....	.....
Total	88.25	89.29	90.08

#### IV. CONCLUSION AND FUTURE WORK

In this paper, we briefly presented a technique for gesture classification in static images using Gabor Filter based features and Bayesian classifier. Our proposed technique shows promising results on our data set. The main point is that the hand segmentation part in our technique plays very important role. If the segmentation is good, the results of gesture classification improve also. The classification accuracy is obtained with the help of Bayesian and Naïve Bayes classifier. In future we intend to extend this work on dynamic gestures as well. Human Computer Interaction and Sign Language Recognition are two of the most prime applications of this research. We also need to compare our method with more state of the art methods and improve it further to incorporate more gestures.

#### ACKNOWLEDGMENT

The author would like to acknowledge the Institute Of Space Technology, especially HEC for providing research facilities for this research.

#### REFERENCES

[1] Ahmad Yahya Dawod, Junaidi Abdullah and Md. Jahangir Alam (2010) "Adaptive Skin Color Model for Hand Segmentation" 2010 International Conference on Computer Applications and Industrial Electronics (ICCAIE 2010), December 5-7, 2010, Kuala Lumpur, Malaysia.

[2] Huang D, Xiang C and Ge SS (2007) "Feature extraction for face recognition using recursive Bayesian linear discriminant". In International symposium on image and signal processing and analysis, Istanbul, Turkey, pp 356–361.

[3] Kwok Wai Wong, kin-man lam and wan-chi Siu (2003) "an efficient color compensation scheme for skin Color segmentation" 0-1803-7161-3/03\$11.00 02003 LEEE.

[4] Wei-Hua, Rewwang and Chun liangtung (2008) "Dynamic Hand Gesture Recognition Using Hierarchical Dynamic Bayesian Networks Through Low-Level Image Processing" Proceedings of The Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 july2008.

[5] Reflex-UL-Zaman Khan and Noor Adnan Ibraheem (2012) "hand gesture recognition: a literature review" International Journal of Artificial Intelligence & Applications (IJAA), Vol.3, No.4, July 2012.

[6] Swapnil D. Badgajar, Gourab Talukdar, Omkar Gondhalekar, Mrs. S.Y. Kulkarni (2014) "Hand Gesture Recognition System," International Journal of Scientific and Research Publications, Volume 4, Issue 2, February 2014 1 ISSN 2250-3153.

[7] Deepika Tewari, Sanjay Kumar Srivastava, (2012) a Visual Recognition of Static Hand Gestures in Indian Sign Language based on Kohonen Self- Organizing Map Algorithm, International Journal of Engineering and Advanced Technology (IJEAT), Vol.2, Dec 2012, pp. 165- 170.

[8] ShrikantKulkarni, Manoj H, Sumam David, VenugopalaMadumbu and Y Senthil Kumar (2011) "Robust Hand Gesture Recognition System Using Motion Templates" 2011 11th International Conference on ITS Telecommunications 978-1-61284-671-2/11/\$26.00.

[9] Wenjun Tan, Chengdong Wu, Shuying Zhao and Shuo Chen (2009) "Hand Extraction Using Geometric Moments Based on Active Skin Color Model" 978-1-4244-4738-1/09/\$25.00 ©2009 IEEE.

[10] Surve Pranjali and Prof. Ubale V.S (2015) "Recognition of Hand Gesture using PCA, Image Hu's Moment and SVM" IPASJ International Journal of Electronics & Communication (IJECE). Volume 3, Issue 4, April 2015 ISSN 2321-5984.

[11] M. A. MONI and A B M Shawkat Ali (2009) "HMM based Hand Gesture Recognition: A Review on Techniques and Approaches" 978-1-4244-4520-2/09/\$25.00 ©2009 IEEE.

[12] Mrsayeshabutalia, Msdivya shah and Dr. R. vdharaskar (2010) "Gesture recognition system" ©2010 International Journal of Computer Applications (0975 – 8887) Volume 1– No. 5.

[13] Doe Hyung Lee and Kwang Seok Hong (2011) "Game Interface using Hand Gesture Recognition" Computer Sciences and Convergence Information Technology (ICCIT), 2010 5th International Conference on Issue Date: Nov. 30 2010-Dec. 2 2010On pages (s): 1092 - 1097Location: Seoul Print ISBN: 978-1-4244-8567-3References Cited: 10INSPEC Accession.

[14] Xiao long Teng, Bin Wu and Weiwei Yu, Chongqing Liu (2005) "A hand gesture recognition system based on local linear embedding" Journal of Visual Languages and Computing 16 (2005) 442–454, www.elsevier.com/locate/jvlc 1045-926X/\$.

[15] Eshed Ohn-Bar and Mohan Manubhai Trivedi (2014), "Hand Gesture Recognition in Real-Time for Automotive Interfaces: A Multimodal Vision-based Approach and Evaluations" To Appear in IEEE Transactions on Intelligent Transportation Systems 2014.

[16] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Kari Pulli, (2015) "Multi-sensor System for Driver's Hand-Gesture Recognition" NVIDIA Research, Santa Clara, California, USA. IEEE International Conference on Automatic Face and Gesture Recognition (FG2015), May 2015.

[17] Hareem Iftikhar, Khurram Khurshid, Fusion of Gabor Filter and Morphological Operators for the Detection of Settlement Zones in Google Earth Satellite Images, IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Nov 2011, Malaysia

[18] Nawazish, Neural Network Based Hand Gesture Recognition, First International Conference on Technology for Helping People with Special Needs, Saudi Arabia, 2013.

# Moon Landing Trajectory Optimization

Ibrahim Mustafa MEHEDI

Center of Excellence in  
Intelligent Engineering Systems (CEIES)  
King Abdulaziz University, Jeddah 21589, Saudi Arabia  
AND  
Electrical and Computer Engineering Department  
King Abdulaziz University, Jeddah 21589, Saudi Arabia  
Email: imehedi@kau.edu.sa

Md. Shofiquel ISLAM

Electrical and Computer Engineering Department  
King Abdulaziz University  
Jeddah 21589,  
Saudi Arabia

**Abstract**—Trajectory optimization is a crucial process during the planning phase of a spacecraft landing mission. Once a trajectory is determined, guidance algorithms are created to guide the vehicle along the given trajectory. Because fuel mass is a major driver of the total vehicle mass, and thus mission cost, the objective of most guidance algorithms is to minimize the required fuel consumption. Most of the existing algorithms are termed as “near-optimal” regarding fuel expenditure. The question arises as to how close to optimal are these guidance algorithms. To answer this question, numerical trajectory optimization techniques are often required. With the emergence of improved processing power and the application of new methods, more direct approaches may be employed to achieve high accuracy without the associated difficulties in computation or pre-existing knowledge of the solution. An example of such an approach is DIDO optimization. This technique is applied in the current research to find these minimum fuel optimal trajectories.

**Keywords**—lunar landing; trajectory optimization; optimization techniques; DIDO optimization

## I. INTRODUCTION

Return to the moon becomes a demanding issue. A lot of scientists and engineers confirmed considerable interests in the past couple of decades [1, 2, 3, 7, 8, 9]. Two-dimensional solution for lunar descent and landing from orbital speed conditions is presented in [4]. The same guidance scheme is improved to 3-dimensional point of view in [6]. More precise 3-Dimensional advanced solution of lunar descent and landing is discussed in [5]. Trajectory optimization capabilities during the Apollo era were severely limited by computing power. An analytic solution for the one-dimensional vertical terminal descent of a lunar soft-landing, based on an application of Pontryagin’s minimum principle, was found by Meditch [10] in 1964. Meditch showed the existence of an optimum thrust program that achieves soft landing under powered descent. Extensive numerical research on the one-dimensional problem was performed by Teng and Kumar [11], using various cost functionals. Their method is based on a time transformation, applied to the calculus of variations. The solution was found numerically, using a quasi-linearization method. In 1971, Shi and Eckstein [12] derived an exact analytic solution for the problem which Teng and Kumar addressed. With the increase in computing power, trajectory optimization techniques, of the type to be discussed in Section 3, have greatly increased the feasibility of generating optimal trajectories with higher complexity and applicability. Recently, Vasile and

Floberghagen [13] applied a Spectral Elements in Time (SET) approach to the lunar soft-landing problem. Within the work, a lunar landing descent from three parking orbit scenarios down to an altitude of 2 [m] above the surface was optimized. The cost function used was based on the square of the control input, which has been noted to be different from the minimum fuel solution [14]. The optimization method used to explore the Moon landing problem in the current research is the Legendre Pseudospectral Method. The Legendre Pseudospectral Method has been applied to a variety of trajectory optimization problems, including problems of ascent guidance [15], satellite formation flying [16], and impulsive orbit transfers [17]. DIDO [18] is used to implement the Legendre Pseudospectral Method in the current research which is discussed in this paper.

The rest of present paper is arranged as follow: The optimization theory has been presented in Section II. The methods of optimization have been explained in Section III. DIDO optimization is introduced in Section IV. The Section V Illustrates the vehicle specification for Moon landing. The Section VI and VII demonstrate the trajectory results and optimized responses with discussion. The conclusion has been given in Section VIII.

## II. OPTIMIZATION THEORY

To maximize or minimize a specified criterion is the task of an optimization problem. This criterion is a function of the parameter which is referred to as a cost function. On this cost function, optimization takes place. Let  $x \in \mathfrak{R}^n$  be a vector of parameters and let  $J(x): \mathfrak{R}^n \rightarrow \mathfrak{R}$  define a cost function. It is desired to find the value  $x^*$  which minimizes  $J$  out of all admissible  $x$ . The global minimum is defined as:

$$J(x^*) \leq J(x) \text{ for all admissible } x \quad (1)$$

A local minimum is defined as:

$$J(x^*) \leq J(x) \text{ for all } x \text{ in the neighborhood of } x^* \quad (2)$$

In a globally convex system, any local optimum found is also the global optimum. On the other hand if the function is non-convex, finding a local optimum does not necessarily imply that it is also the global optimum solution. Fig. 1 illustrates the concept of convexity vs. non-convexity and local vs. global optimums for an arbitrary one-dimensional function  $f(x)$ .

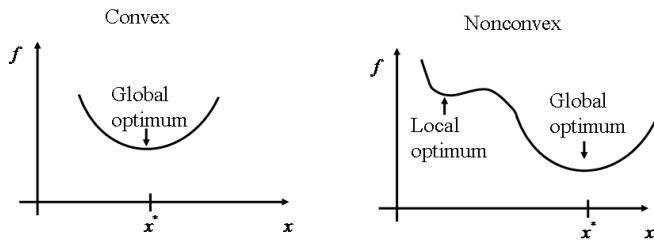


Fig. 1. One-dimensional convexity

Only one optimal solution which is the global optimum is available in the example shown for convexity. Two optimal, local and global optimum solutions exist in the non-convex example. Most trajectory optimization problems are non-convex, and therefore, only local optimums are readily found.

A parameter optimization problem that has a linear cost function, linear constraints, and has only real values is known as a linear programming problem. The term is only used for problems which have real values; otherwise, it falls into the category of integer or mixed-integer programming. If the problem has only real values but includes a nonlinear cost function or nonlinear constraints, it is referred to as a nonlinear programming (NLP) problem.

From basic calculus, recall that a local minimum of a function, which is the function of one variable (i.e.,  $f(x)$ ), can be found by locating a point where the first derivative of the function on the variable equals zero and the second derivative is positive. This notion is expanded to higher dimensions with gradients and Hessians. A gradient is a vector consisting of the first-order partial derivatives of a function on each variable. The Hessian is a matrix consisting of the second-order partial derivatives of the function. For example, the following conditions are satisfied at a local minimum for the cost function  $J(x)$ :

$$\frac{\partial J}{\partial x} = 0 \quad (3)$$

$$\frac{\partial^2 J}{\partial x^2} \geq 0 \quad (4)$$

One could attempt to determine the minimum of the cost function analytically, but it may be difficult or impossible. Therefore, iterative techniques are used to locate the minimum by searching over the region of admissible  $x$ . Newton's method, the most common technique, uses the gradient and Hessian information at the current location to determine a search direction. There are numerous other iterative techniques used, most being based on the principle of Newton's method.

### A. Constrained Optimization

Constrained optimization problem can be either equality constraints or inequality constraints. While a function, or functions, of the parameters, equal to a specific value, then it is added to the optimization problem as an equality constraints. The parameters can vary, but relationships between the parameters remain fixed. An equality constraint has the form:

$$f(x) = 0 \quad (5)$$

where  $f: \mathbb{R}^n \rightarrow \mathbb{R}^q$

Formulating the augmented cost function, the most convenient method of solving equality constrained

optimization problem, which is a combination of the cost function and the constraints with multipliers.

$$J' = J(x) + \lambda^T f(x) \quad (6)$$

To find the minimum, the gradient of  $J'$  is taken on  $x$  and  $\lambda$  and set to zero. The first-order necessary conditions for optimality become:

$$\frac{\partial J'}{\partial x} = \frac{\partial J}{\partial x} + \frac{\partial f}{\partial x} \lambda = 0 \quad (7)$$

$$\frac{\partial J'}{\partial \lambda} = f(x) = 0 \quad (8)$$

On the other hand, an inequality constraint has the form:

$$g(x) \leq 0 \quad (9)$$

where  $g: \mathbb{R}^n \rightarrow \mathbb{R}^q$

The constraints are adjoined to the cost function in a similar manner as the equality constraints, but with the multipliers  $\mu$ . The augmented cost function is now defined as follows:

$$J' = J(x) + \lambda^T f(x) + \mu^T g(x) \quad (10)$$

Inequality constraints can be categorized into either 'active' or 'inactive' classes. The vector  $g(x)$  can be written as seen in (11), with  $g_i(x)$  for ( $i = 1..p$ ) representing individual components of the  $p$ -dimensional vector. An active constraint is when a component of the constraint vector equals zero at the optimum solution, (i.e.,  $g_i(x^*) = 0$ ). The optimization problem is bounded by these constraints. Inactive constraints are constraints where  $g_i(x^*) < 0$  and these constraints do not affect the optimal solution. The equation  $\mu^T g = 0$ , which is known as complementary slackness, ensures that the individual inequality constraints are either active or do not affect the solution.

$$g = \begin{bmatrix} g_1(x) \\ \vdots \\ g_p(x) \end{bmatrix} \quad (11)$$

Karush-Kuhn-Tucker (KKT) conditions are named for the first-order necessary optimality conditions. These are given by (8) and (9), and the following:

$$\frac{\partial J'}{\partial x} = 0 \quad (12)$$

$$\mu \geq 0 \quad (13)$$

$$\mu^T g = 0 \quad (14)$$

### B. Concept of optimal control

A particular category of optimization problems which comprise dynamical constraints that vary with time is recognized as functional optimization problems, where the term "functional" is used to denote a function of a function. Functional optimization problems that have an input, or control, to be determined are identified as an optimal control problems. Optimal control problems have an extensive diversity of applications, together with the field of trajectory optimization.

Let  $x \in \mathbb{R}^n$  be the state of a continuous system where  $t \in \mathbb{R}$  is time. Furthermore, let  $u(t) \in \mathbb{R}^n$  be the control of input. Lastly, let the dynamical constraints that govern the change of  $x(t)$  with respect to time be given as:



$$\dot{x} = f(x(t), u(t), t) \quad (15)$$

To search for the function  $\mathbf{u}(t)$  that minimize the cost functional,  $J$  is a general optimal control problem, with the condition that the constraints imposed on the problem. This problem should include the dynamical constraints. It is normally desired to either minimize a functional of the state and control over the entire time span or the final value of a criterion. The cost functional is therefore composed of two parts: a terminal cost  $\Phi: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$  (commonly referred to as the Mayer cost) and an integrated cost  $L: \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}$  (known as the Lagrange cost). The Bolza form of the cost functional, as seen in (16), is a combination of the terminal and integrated costs.

$$J = \Phi(x(t_f), t_f) + \int_{t_0}^{t_f} L(x(t), u(t)) dt \quad (16)$$

The initial and final boundary conditions that the system must satisfy are given by (17) and (18), respectively.

$$\varphi_0(x(t_0), t_0) = 0 \quad (17)$$

$$\varphi_f(x(t_f), t_f) = 0 \quad (18)$$

Where  $\varphi_0: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^{q_0}$  and  $\varphi_f: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^{q_f}$ .

Augmented cost functional  $J'$  is formulated to combine all of this information. A vector of Lagrange multipliers is called the costate. This costate adjoined by the dynamical and terminal constraints to the cost functional. Here,  $\lambda(t) \in \mathbb{R}^n$  is Lagrange multipliers, and  $v \in \mathbb{R}^{q_f}$  is terminal constraint multipliers. The equation given below is similar in Reference [19].

$$J' = \Phi(x(t_f), t_f) + v^T \varphi(x(t_f), t_f) + \int_{t_0}^{t_f} [L(x(t), u(t), t) + \lambda(t)^T (f(x(t), u(t), t) - \dot{x})] dt \quad (19)$$

Hamiltonian and end point functional are defined in (20) and (21), respectively to simplify the augmented cost functional.  $J'$  is simplified to (22). The functional dependencies have been omitted for clarity.

$$H = L + \lambda^T f \quad (20)$$

$$G = \Phi + v^T \varphi \quad (21)$$

$$J' = G + \int_{t_0}^{t_f} [H - \lambda^T \dot{x}] dt \quad (22)$$

It is determined a stationary point of the augmented cost functional using calculus of variations. Using the previous concept, the variation of  $J'$  is taken and appropriate functions are set to zero. The resulting necessary conditions for a local minimum for the free final time problem are derived by Hull [20] as:

$$\dot{x} = \frac{\partial H}{\partial \lambda} = f(x, u, t) \quad (23)$$

$$\dot{\lambda} = -\frac{\partial H}{\partial x} \quad (24)$$

$$0 = \frac{\partial H}{\partial u} \quad (25)$$

$$\varphi_0(x(t_0), t_0) = 0 \quad (26)$$

$$\varphi_f(x(t_f), t_f) = 0 \quad (27)$$

$$\lambda^T(t_f) = \left(\frac{\partial G}{\partial x}\right)_{t=t_f} \quad (28)$$

$$(H)_{t=t_f} = -\left(\frac{\partial G}{\partial x}\right)_{t=t_f} \quad (29)$$

### III. OPTIMIZATION METHODS

Optimality conditions described in the previous are to be solved to find a stationary point. An analytical solution is almost impossible for most non-linear optimization problems. Therefore, computer-aided numerical methods are the alternate way in finding the optimal solution. There are numerous numerical methods that have been formulated to solve optimal control problems, and it is not in the scope of this thesis to explore them all. The method used in this research is the Legendre Pseudospectral Method, which is a direct transcription method that uses a spectral technique. The meanings of 'direct', 'transcription', and 'spectral' will be discussed below.

#### A. Direct Methods

Computer aided numerical methods can be divided into two distinct categories: direct and indirect. An indirect method uses information from the costate differential, the maximum principle, and the boundary conditions shown in (26 - 29), to find the optimal solution. To use this method, an estimate of the costate is required a priori, which may pose a problem since the costate does not usually have physical significance.

The direct method is commonly used to optimize directly the cost function such as (16). The method searches for the feasible region for a minimum of the cost function starting from an initial guess of the state and control. It is reported as a local minimum solution because it is impossible to search the entire feasible region. If it can be proven that the problem is convex, then local optimality implies global optimality. Being familiar with the dynamics of the problem, it can provide an initial guess of the state and control rather than costate. In some cases, a simple propagation of the state from the desired initial conditions with no control input is sufficient. However, if the problem is highly non-convex, a good initial guess may be crucial in finding the correct local optimal solution.

#### B. Direct Transcription Methods

The direct method implies some different ways such as direct shooting methods, direct transcription, etc. A direct transcription method is used in this research and described as follows.

Transformation of continuous time domain system into discrete time domain system is an important issue in real-time optimal control problem to be implemented on a computer. The locations in time at which the problem is discretized are referred to as 'nodes', and can be uniformly or non-uniformly distributed in the time domain. At each node, the discrete system represents the continuous system, and links must be made between the nodes to represent the dynamics of the original continuous system. A transcription method is used to transform the continuous system into the discrete problem. Most optimal control problems include either a nonlinear cost function or nonlinear constraints (which may include nonlinear dynamical constraints) that are only functions of real variables. As a result, the transcribed problem is an NLP.

Two consecutive tasks are performed by this transcription method. One is to transform the problem from a continuous system to a discrete system, and another is to link the nodes together in a way that represents the dynamics of the original

problem. Representing the dynamics of a continuous system is done by different ways such as the Euler method, the Runge-Kutta method, and spectral methods. The method used in this research is a spectral method, which fits globally orthogonal polynomials to the discrete data over the entire time span. The Legendre Pseudospectral Method uses a special class of orthogonal polynomials, known as Legendre Polynomials. The interior nodes are placed at the roots of the Legendre polynomial derivative, known as the Legendre-Gauss-Lobatto (LGL) points, which provides higher accuracy in the results.

#### IV. DIDO OPTIMIZATION

In the past, it was an extremely difficult task to analyze fully a complex system involving numerous dependent variables. It was an indirect and approximate method to allow solutions to be found. But now computers have improved exponentially in processing speed and ability, allowing complex computations to be completed quickly and efficiently without imposing a significant drain on resources. One example of this improvement is in the field of optimization analysis. With the emergence of such improved processing power and the application of new methods, more direct approaches may be employed to achieve high accuracy without the associated difficulties in computation or pre-existing knowledge of the solution. An example of such an approach is DIDO optimization. The method is named for Queen Dido of Carthage (circa 850BC) who was the first person known to have solved a dynamic optimization problem.

DIDO relies on the Legendre Pseudospectral Method, which has been developed and employed primarily in fluid flow modeling. This method employs Legendre Polynomials to create an approximation of variables over multiple nodes as opposed to the use of a fixed order polynomial. This allows, despite discontinuities in the governing equations, a solution to be attained with high accuracy which also satisfies the imposed optimization criterion, where most direct methods do not [21]. The result is a method of solving a complex dynamic optimization problem without a priori knowledge of the solution or incredibly complex analytic computations. This is exactly the type of method required to optimize a problem as complex as an examination of the trade space for lunar landing.

The DIDO optimization code requires the development of five different files to specify the problem accurately. Once the problem is defined, DIDO optimizes a solution using the restrictions found in the problem definition and on the cost function as previously discussed. These files include the Cost Function File, Dynamics-Function File, Events File, Path File and Problem File [22].

#### V. VEHICLE SPECIFICATION

The vehicle is assumed to be axially symmetric with a thrust to mass ratio that varies between 4.0 [N/kg] and 10.0 [N/kg]. It assumed that the vehicle has one throttle-able main engine, which is fixed to the vehicle and does not gimbal. For a hypothetical initial mass of 1800 [kg] (which is representative of an unmanned vehicle), a value of 8000 [N] was chosen for the engine's maximum thrust limit,  $F_{max}$ . The engine's exhaust velocity,  $V_{ex}$ , was chosen as 3500 [m/s], which corresponds to a specific impulse of 358 s.

In comparison, the Surveyor vehicles had final landed masses varying from 1431[kg] to 1486 [kg] [23, 24]. Each lander had a propulsion system which consisted of one main solid propellant retro-fire engine and three throtttable liquid propellant vernier engines. The vernier engines, which were used to remove the final 100 [m/s] of velocity and also for attitude control during the retro-fire, were capable of providing 133 [N] to 463 [N] of thrust each. They had a specific impulse which varied from 273 [s] at minimum thrust to 287 [s] at maximum thrust [25].

#### VI. TRAJECTORY RESULTS

This section explores the results obtained during trajectory optimization of the Moon landing problem. The equations of motion derived in References [4, 5, 6,] were coded in DIDO. In this section, all work is limited to planar motion, but can easily be extended to three-dimensions. The convexity of this problem was not determined; therefore, all results are reported as local optimal solutions.

An initial parking orbit 100 [km] was chosen for the analysis throughout this simulation. The objective of the current research is to minimize fuel usage while meeting specified operational constraints. The cost function was chosen to be the final mass of the vehicle because this is the most direct measurement of fuel usage, as seen in (30).

$$J_{min} = -m(t_f) \quad (30)$$

The main criterion is to land with almost zero velocity at the surface while using bounded thrust.

##### A. Boundary Conditions

The vehicle parameters, as described in Section 5, are listed along with the initial and end conditions in Table 1 and 2.

TABLE I. INITIAL CONDITIONS AND VEHICLE PARAMETERS

Parameters	Value	Unit
Maximum available thrust	8000	[N]
Specific Impulse	358	[s]
Initial altitude	100	[km]
Initial down range distance	0	[km]
Initial velocity vector pitch angle	90	[deg]

TABLE II. END CONDITIONS OF POWERED DESCENT CONDITIONS

Parameters	Value	Unit
Final altitude	1	[km]
Final down range distance	425	[km]
Final velocity vector pitch angle	0.1	[deg]
Lunar gravitation acceleration	1.623	[m/s <sup>2</sup> ]

#### VII. RESPONSES

Initially, it is assumed that the expected landing point is about 425 [km] far. Other initial and end conditions are dictated in Table 1 and 2. The physical path of the trajectory is shown in Fig. 2. The path solved for by DIDO and depicted in the Fig. is the most efficient in terms of fuel consumption, and thus is the optimal solution regarding this analysis.

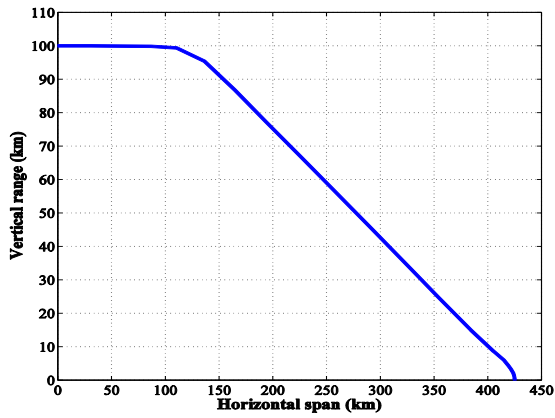


Fig. 2. Optimized lunar landing trajectory

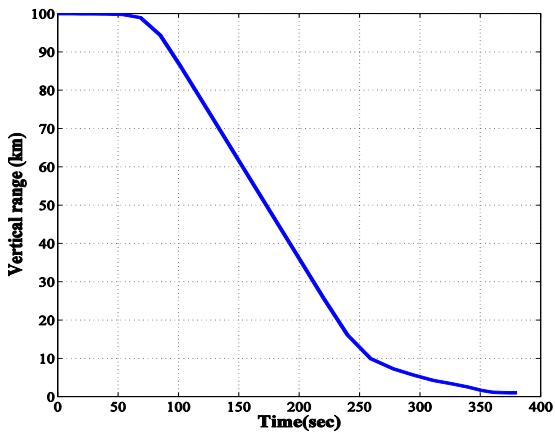


Fig. 3. Optimized lunar landing trajectory: Vertical range

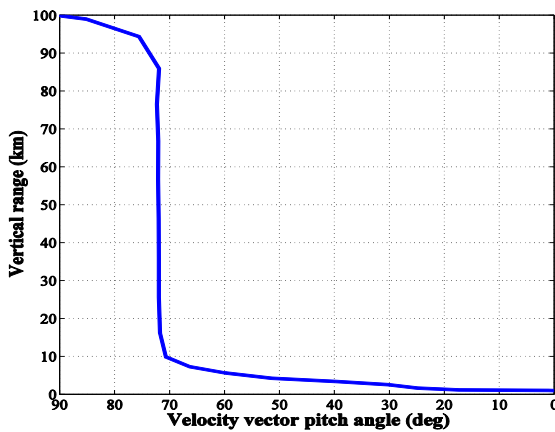


Fig. 4. Optimized lunar landing trajectory: Vertical range as a function of velocity

To instill a better understanding of the trajectory, the vertical and horizontal positions relative to time are displayed in Figs 3 and 5. Further, trajectory profiles regarding vertical and horizontal positions as a function of velocity vector pitch angle are shown in Figs 4 and 6. Fig. 3 illustrates the parametric shape of the burn that takes place, indicating

features of perfect fuel optimization. The burn is timed precisely, in this case, to minimize the amount of fuel consumed.

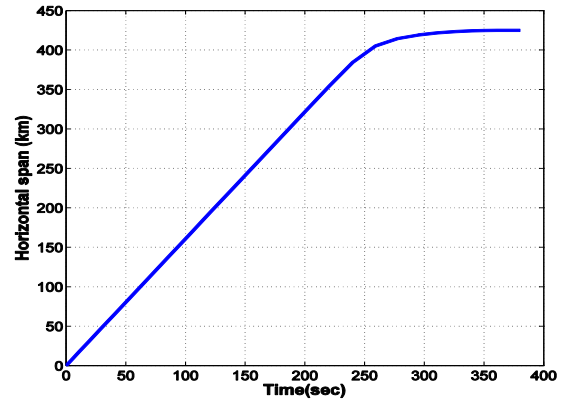


Fig. 5. Optimized lunar landing trajectory: Horizontal span

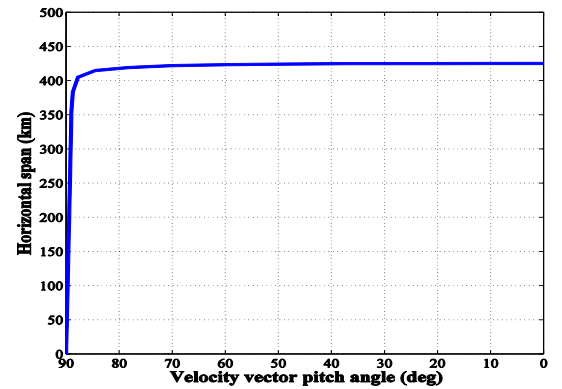


Fig. 6. Optimized lunar landing trajectory: Horizontal span as a function velocity vector pitch angle

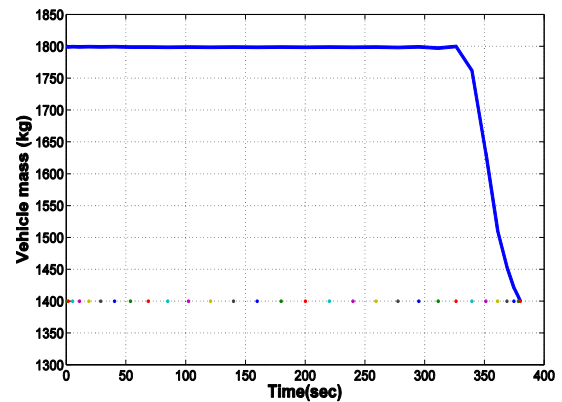


Fig. 7. Optimized lunar landing trajectory: Vehicle mass

Fig. 5 demonstrates the nearly constant progression in the horizontal plane. Another parameter of interest is that which drives the optimization, mass, and is shown in Fig. 7. It is observed that the 400 [kg] fuel was used to minimize the amount of addition fuel being carried. The intent is that this approximately simulates the eventual fuel budget of the lunar

landing vehicle though safety reserves and other factors will affect the final value. When compared to Fig. 3 it can be seen that the main usage of fuel occurs during the vertical braking burns, as would be expected due to a large amount of thrust required during these periods. Movement in the horizontal plane causes this line to continuously decrease, but at a much slower rate, due to the relatively small amount of thrust in this direction.

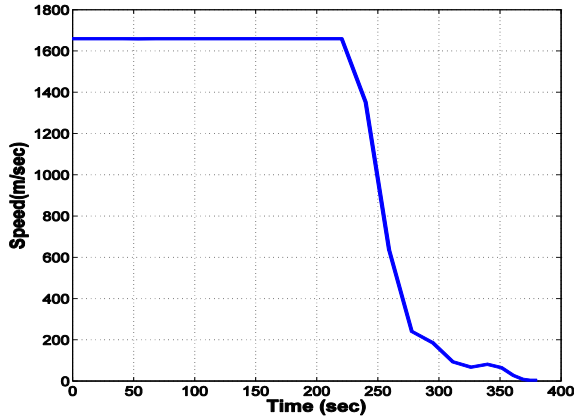


Fig. 8. Optimized lunar landing trajectory: Vehicle speed

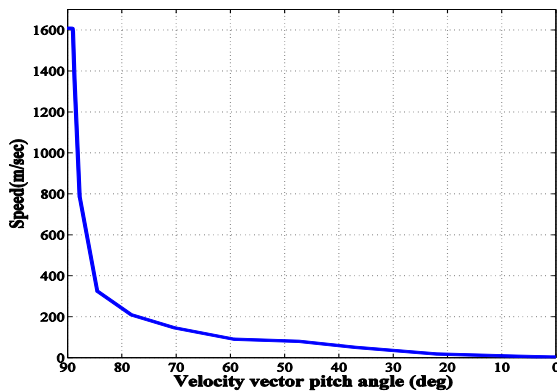


Fig. 9. Optimized lunar landing trajectory: Vehicle speed as a function of velocity vector pitch angle

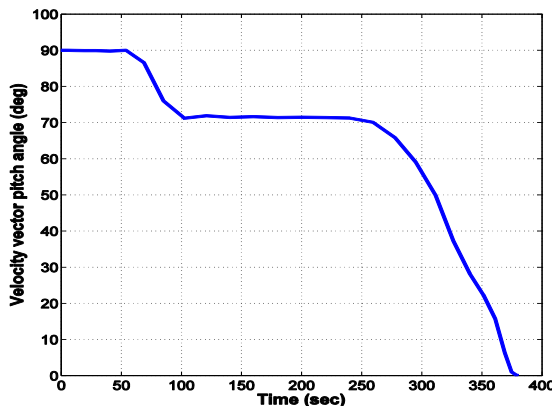


Fig. 10. Optimized lunar landing trajectory: Velocity vector pitch angle

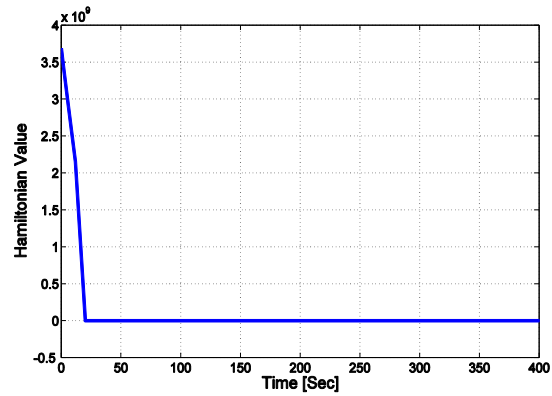


Fig. 11. Hamiltonian evaluation for optimality verification

Details of trajectory, including vehicle speed against time and velocity vector pitch angle are shown in Fig. 8 and 9, respectively. It is important to note the difference in y-axis scales on the vehicle speed plots. This trajectory is largely horizontal, having a maximum velocity of 1609 [m/s] which decreases rapidly once the final braking burn commences. Again, the velocity vector pitch angle is plotted against time in Fig. 10.

## VIII. CONCLUSION

According to the Minimum Principle, given a candidate optimal solution, there exists a collection of dual functions and variables such that certain statements are true. Depending upon the problem these statements are examined as a means to test the optimality of the DIDO run. From the Minimum Principle, it must satisfy the Hamiltonian Minimization Condition [26, 27], Therefore, the current investigation with the help of DIDO includes the results of Hamiltonian. This is plotted in Fig. 11. The Hamiltonian should be constant over the entire time span. Looking at the Fig. 11, the estimate of the Hamiltonian is fairly small, but oscillation is present. A large fluctuation is seen where the throttle switch occurs, because of the initialization in the control during descent start. Thus, the computer simulation has indeed found the exact optimal solution at a very high precision. Using this direct approach like DIDO optimization helped to achieve high accuracy without the associated difficulties in computation or pre-existing knowledge of the solution.

## ACKNOWLEDGMENT

This article was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah. Therefore, the authors acknowledge with thanks DSR financial support

## REFERENCES

- [1] R. Cheng, "Lunar terminal guidance In Lunar Missions and Exploration," edited by C. T. Leondes and R. W. Vance, Univ. of California Engineering and Physical Sciences Extension Series, Wiley, New York, pp. 308–355, 1964.
- [2] R. Cheng, C. Meredith and D. Conrad, "Design considerations for survivor guidance," Journal of Spacecraft and Rockets, Vol. 3, Issue 11, p. 1569–1576, 1966.
- [3] R. Klumpp, "Apollo guidance, navigation, and control: Apollo lunar-descent guidance." Technical report, MIT Charles Stark Draper Laboratory, 1971.

- [4] I. M. Mehedi and T. Kubota, "Advanced guidance scheme for lunar descent and landing from orbital speed conditions," *Transaction of Japan Society for Aeronautical and Space Sciences*, Vol. 54, Issue 184, pp. 98–105, 2011.
- [5] I. M. Mehedi, T. Kubota and U. M. Al-Saggaf UM, "3- Dimensional advanced solution of lunar descent and landing," *Journal of Vibroengineering*, Vol. 17, Issue 1, pp.527–543, 2015.
- [6] I. M. Mehedi, T. Kubota and U. M. Al-Saggaf UM, "3- Dimensional Analytical solution for lunar descent scheme," *Iranian Journal of Science and Technology Transactions of Electrical Engineering*, Vol. 39, Issue E1, pp. 67–77, 2015.
- [7] R. Sostaric, "Lunar descent reference trajectory," Technical report, NASA/JSC, 2006.
- [8] S. Ueno and Y. Yamaguchi, "Near-minimum guidance law of a lunar landing module," 14th IFAC Symposium on Automatic Control in Aerospace, pp. 377–382, 1998.
- [9] L. Xing-Long, D. Gaung-Ren and T. Kok-Lay, "Optimal soft landing control for moon lander," *Automatica*, Vol. 44, pp. 1097–1103, 2008.
- [10] J. S. Meditch, "On the Problem of Optimal Thrust Programming for a Lunar Soft Landing," *IEEE Transactions on Automatic Control*, Vol. AC-9, No. 4, pp. 484, 1964.
- [11] L. Teng and K. S. P. Kumar, "Optimum Control for Lunar Soft Landing," *Astronautica, Acta*, Vol. 13, No. 5-6, pp. 545, 1968.
- [12] Y. Y. Shi and M. C. Eckstein "An Exact Solution For Optimum Controlled Soft Lunar Landing," Vol. 16, No. 1, pp. 18, 1971.
- [13] M. Vasile and R. Floberghagen, "Optimal Trajectories for Lunar Landing Missions," *AAS/GSFC International Symposium, Advances in the Astronautical Sciences*, Greenbelt, MD, pp. 281-295, 1998.
- [14] I. M. Ross, "How to Find Minimum-Fuel Controllers," *AIAA Guidance, Navigation, and Control Conference and Exhibit*, American Institute of Aeronautics and Astronautics, Inc., Providence, RI, pp. 1-10, AIAA Paper 2004-5346, 2004.
- [15] Rea, J. R. A Legendre Pseudospectral Method for Rapid Optimization of Launch Vehicle Trajectories, Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, May 2001.
- [16] S. I. Infeld, S. B. Josselyn and W. Murray, "Design and Control of Libration Point Spacecraft Formations," *AIAA Guidance, Navigation, and Control Conference and Exhibit*, Providence, RI, pp. 1–14, AIAA Paper 2004-4786, 2004.
- [17] S. A. Stanton, "Optimal Orbital Transfer using a Legendre Pseudospectral Method," Master's thesis, Massachusetts Institute of Technology, 2003.
- [18] I. M. Ross, "User's Manual for DIDO (Ver. PR.13): A MATLAB Application Package for Solving Optimal Control Problems," Technical Report 04-01.0, Naval Postgraduate School, Monterey, CA, February 2004.
- [19] A. Bryson and Y. C. Ho, "Applied Optimal Control-Optimization, Estimation and Control," Hemisphere Publishing Company, 1975.
- [20] D. Hull, "Optimal Control Theory for Applications," *Mechanical Engineering Series*, Springer-Verlag New York, Inc., New York, NY, 2003.
- [21] M. Ross, and F. A. Fahroo, "Perspective on Methods for Trajectory Optimization," *AIAA/AAS Astroynamics Specialist Conf.*, Monterey, 2002.
- [22] H. B. Robert and A. C. Antoulas, "Nonlinear Approach to Aircraft Tracking Problem," *Journal of Guidance, Control, and Dynamics*, 17(5), pp. 1124-1130, October 1994.
- [23] Surveyor Project Final Report, Part 2. Project "Description and Performance," Tech. Rep. JPL-TR-32-1265, NASA, Pasadena, CA, July 1969.
- [24] "Surveyor I Mission Report, Part 1. Mission Description and Performance," Tech. Rep. JPL-TR-32-1023, NASA, Pasadena, CA, August 1966.
- [25] M. E. Ellion, H. DiCristina, A. R. Maffei, and A. Briglio, "Development of the Surveyor Vernier Propulsion System (VPS)," *AIAA Second Propulsion Joint Specialist Conference*, Colorado Springs, CO, 1966.
- [26] R. Vinter, "Optimal Control," BirkhAuser, Boston, MA 2000
- [27] I. M. Ross, "Control and Optimization: An Introduction to its Principles and Applications," Electronic Edition, Naval Postgraduate School, Monterey, CA, October 2006.

# Evaluation of Navigational Aspects of Moodle

Raheela Arshad, Awais Majeed\*  
Department of Software Engineering  
Bahria University  
Islamabad, Pakistan

Hammad Afzal  
Department of Computer Software  
Engineering  
National University of Sciences &  
Technology (NUST)  
Islamabad, Pakistan

Muhammad Muzammal, Arif ur  
Rahman  
Department of Computer Science  
Bahria University  
Islamabad, Pakistan

**Abstract**—Learning Management System (LMS) is an effective platform for communication and collaboration among teachers and students to enhance learning. These LMSs are now widely used in both conventional and virtual and distance learning paradigms. These LMSs have various limitations as identified in the existing literature, including poor learning content, use of appropriate technology and usability issues. Poor usability leads to the distraction of users. Literature covers many aspects of usability evaluation of LMS. However, there is less focus on navigational issues. Poor navigational can lead to disorientation and cognitive overload of the users of any Web application. For this reason, we have proposed a navigational evaluation framework to evaluate the navigational structure of the LMS. We have applied this framework to evaluate the navigational structure of Moodle. We conducted a survey among students and teachers of two leading universities in Pakistan, where Moodle is in use. This work summarizes the survey results and proposes guidelines to improve the usability of Moodle based on the feedback received from its users.

**Keywords**—*e-Learning; Navigational Evaluation Framework; Learning management system (LMS); Moodle; Usability*

## I. INTRODUCTION

Rapid increase in technology provided educational institutions the opportunity to use Internet as a communication source for instructions. The notions used in technology-based learning are now becoming significant with e-learning and have become a vital medium for educational organizations for an efficient provision of learning material [1]. For e-learning environments, it is important that they should be designed and evaluated in an academically efficacious manner by taking into consideration both pedagogical and usability concerns [2]. A reliable foundation of the e-learning platform is learning management system (LMS) that complies with educational standards and best practices endorsed by experienced educationists [3]. However, despite the prevalent use of LMSs, the critical examination of their usability is a relatively new area of research [4]. Studies [2] [5] showed that although there are many other reasons of high distraction in e-learning platforms such as inappropriate use of contents and technology, but the key issue is the poor usability of such platforms. Therefore, usability of such applications should be very high. When speaking of usability of LMS, we navigation is a vital factor to be considered and required to be evaluated. A good navigation structure in LMS helps users accomplishing learning activities without spending too much time to learn how to go to the next page [6].

Moreover, new Web design methodologies treat navigation as a separate concern [7]. Still there is no framework for evaluating the effectiveness of the navigation structure of the LMS. Therefore, in this research, we proposed a framework for evaluation of the navigational structure of the LMS. Moodle is selected for the evaluation and through a survey we received feedback on the evaluation framework from users of Moodle in order to validate the factors of the framework as well as to identify the problems in navigation design of Moodle.

Moodle (Modular Object-Oriented Dynamic Learning Environment) is one of the most widely used open source LMS intended to design complete educational principles to support instructors in creating effective online learning groups. Martin Dougiamas founded Moodle as part of his PhD thesis in 1999 and in 2002 it was released to the public. Design of Moodle is based on constructivism and pedagogical philosophies. Pedagogy is concerned with the strategies of instruction or how best to teach; and constructivism provides an ability for users to share their ideas and experiences to one another [8].

The research questions and hypotheses addressed in this research are:

- 1) *What are the navigational factors that are important in the context of LMS?*
- 2) *How efficient is the navigation of Moodle based on the identified factors?*
- 3) *Is there a difference between responses received from two universities?*

**H1:** *The correlation of ease of navigation experience with satisfaction about navigation design and efficiency of performing the task is significant.*

**H2:** *The correlation of seven navigation factors is significant with ease of navigation and satisfaction about navigation design.*

The assessment is conducted in two universities in Pakistan using the survey technique. The rest of this study is divided into 5 sections. Section II presents related work. Section III describes the research methodology used in this research. Section IV describes important navigation factors of the proposed framework. Section V provides a detailed insight on results and findings. Finally, section VI presents conclusions and future work.



## II. RELATED WORK

### A. Usability in the context of LMS

In case of e-learning platforms, it is not just enough to provide users with such systems that contain extensive features with minimum or no cost at all. Research on Human Computer Interaction (HCI) suggests that the major issue in technology is usability and how easy to use a system is. Usability is a quality attribute as indicated by Jakob Nielsen, a usability expert who brought the concept of Web usability [9]. According to ISO usability is "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use." [10]. Usability of LMS significantly affects the learning, since students' interaction with these systems should result in a successful achievement of learning goals rather to complete a task successfully. Consequently, the interface of LMSs should not distract learners from accomplishing their learning objectives. When these systems are not easy to use, users might spend more time trying to understand how to start interacting with the system, instead of engaging with the actual learning material. Inversini *et al.* [11] indicated that, in terms of teaching and learning, a usable LMS significantly reduces teacher time spent in setting up and managing the course contents and improves the students learning performance as well as experience. Thus, the LMS should have a simple interface with which teachers and students do not need to be worried about difficult technology and can put concentration on learning contents. Usability comprises many factors; among these navigation is one of the major factors that influence the usability. Nielsen [10] indicated that intuitively organized and easy to navigate Web pages, help users find the required information with ease.

### B. Importance of ease of navigation in LMS

Easy navigation is important in any LMS that enhances the learning activities. Users especially students are from different educational backgrounds and some of them are not proficient in using computers and computer applications. LMS is a Web-based system, and therefore, it can become a barrier to learners who are less proficient with computers. To overcome this obstacle, the LMS should be designed in such a way that it is user friendly and has a well-managed interface with good navigation to help users understand the content quickly [12]. Hence, when designing LMS, ease of navigation should be the top priority because if the course is easy to navigate, the learners will have a better learning experience even if they have no or minimal computer skills. Fluid movement within the LMS, allows users to freely explore entire LMS whereas, poorly designed navigation might hamper user interactions and reduce productivity [13].

Good navigation contributes in self-directed learning and facilitates the progress throughout the learning process. Alternatively, difficult navigation creates difficulty for learners to achieve learning outcomes successfully [14]. Also, it is frustrating when users get disoriented while navigating through the LMS. A well-organized LMS enables users to focus on the desired goals.

Without a good navigation provision, students are not able to achieve learning goals because they are not able to know

what page they need to go to obtain desired learning materials [6]. Free navigation through LMS permits the learners to create their specific knowledge structures and it is a critical factor in assisting interactivity among them [15]. If LMS does not permit learners to navigate efficiently, then this can become a hurdle for them to be involved as active participants in the learning process. Thus, navigation is an essential part of any LMSs that enhances the learning activities by delivering the learning materials effectively to students.

### C. Moodle's usability evaluation

In the literature we came across regarding usability evaluation of Moodle, we analysed that the usability of Moodle has been evaluated mostly based on Nielson's heuristics. Further, we investigated some characteristics of usability have been explored such as efficiency, ease of use, effectiveness, and learnability, and few others. We found navigational characteristic has not been explored as a broader usability characteristic. Author in [16] shows the usability of Moodle's registration module and assignment submission module by utilizing the DECIDE framework. Additionally, another similar study [17] evaluated the usability of Moodle different modules by employing questionnaires, task driven and heuristic evaluation techniques and generated qualitative and quantitative data. Beside these data, expert opinions were also given to provide important recommendations for all users to create a more user friendly Moodle. Ivanović *et al.* [18] explored usability of Moodle in order to reflect the teachers' and students' opinions.

User satisfaction is a factor on which usability may exert influence. Also, to achieve learning goals in e-learning platforms, it is a very crucial aspect to be considered. Thus, in [19] user satisfaction was more focused during usability analysis of Moodle in which students' perceptions were more motivated. In this study navigation acquired very little attention. In another study [20], efficiency and ease of learning of Moodle were analysed in order to reflect the usability. Padayachee *et al.* [21] conducted a usability comparison based on Nielsen's heuristics for three LMSs including Moodle, DOKEOS, and ATutor. This study explored issues against each heuristic. In this study heuristic "Flexibility and efficiency of use" covered a navigation aspect related to navigation jumps; for quickly performing desired actions. Another similar study [22] conducted a comparative usability evaluation among Moodle, Sakai, and dotLRN. This evaluation was based on Nielsen's heuristics in which five usability experts followed a task-based approach. Strengths and weaknesses against each heuristic were explored in relation to each LMS. The study presented in [23], is an enhancement of previous one, but in this study pattern-based usability analysis was conducted. Patterns were based on Nielson's heuristics, in which authors selected one pattern against each heuristic. The study presented in [24] also used Nielsen's usability heuristics to compare the usability of three LMSs i.e. WebCT, Sakai, and Moodle with perspectives of students. Another study [25] performed usability comparison between two LMSs: Blackboard and Moodle from the viewpoint of students and faculty. The main focus of this study was on efficiency, ease of use, desired functionality, and user interface. The authors of this study concluded that

Moodle is more efficient and effective than Blackboard LMS. In [26], usability evaluation of Moodle from teachers' and students' perspective was conducted. This evaluation used five usability aspects; efficiency, memorability, errors, satisfaction, and learnability suggested by Nielsen. The previous study is similar to the one presented in [27] that used two additional usability characteristics; operationability and attractiveness, to assess the usability of Moodle. Quality is an important aspect that affects usability as indicated in [28] thus, Oztekin *et al.* [29] proposed a new approach *UseLearn* for usability evaluation of e-learning platforms. It incorporates both usability evaluation perspective and quality in e-learning platforms. In this approach navigation achieved a little attention to evaluate usability of Moodle.

Numerous research studies have been conducted on Moodle's usability evaluation, with very less focus on the navigational evaluation of Moodle. Therefore, exploring the effectiveness of the navigational aspects of Moodle in the context of usability becomes appealing. Moreover, guidelines for the development of navigational structure of the LMS is also missing.

### III. RESEARCH DESIGN

We first propose a navigational evaluation framework for LMS. After that a survey is conducted in order to get the feedback on the evaluation framework from users of Moodle. For this purpose both quantitative and qualitative research methods were used.

Our research is twofold in which two different types of questionnaires, first generic questionnaire and second tasks specific questionnaire, were constructed. Generic questionnaire was comprising questions based on factors of proposed evaluation framework, whereas task based questionnaire contained of questions regarding navigation of the most frequently performed tasks by both teachers and students.

Participants of this study were selected from the two leading universities of Pakistan. Participants were divided into two different groups: students and teachers. In case of selecting sample of students, we conducted our survey on Bachelor's level students' from both universities in which IT students were selected. Reason for selecting the IT students was that they had similar knowledge of using computer.

### IV. NAVIGATIONAL EVALUATION FRAMEWORK FOR LMS

In the context of LMS, the navigation has a significant impact on learning and teaching activities. Thus, it is important that the LMS should have an efficient navigation structure. In addition, there should be a framework against which usability experts and system analysts can evaluate the navigational structure of LMS.

Fig. 1 shows the proposed framework, underlining the most important aspects to be considered for evaluating the navigation structure of the LMS. The framework consists of seven major navigation factors and each major factor further comprises sub factors. The major factors of the framework are

links, navigation menu, shortcut facility, services, navigation aids, consistency, and adaptive navigation.

*Links* category defines the main tool of hypertext design where various nodes (web pages, documents and multimedia) are linked together through hyperlinks. It provides the fundamental navigation mechanism to move around the web site and access various functionalities and perform tasks. Links can be text-based, navigational buttons or in the form of icons/images.

*Menus* are helpful to categorize the hyperlinks for easy navigation. Placement of these menus is significant in the usability of any Web application. Menu position can be vertical or horizontal. Moreover, menu structure in terms of its depth and breadth affects the user's ability to reach a particular page using the menus.

*Shortcuts* increase the work performance of users by making them able to quickly access the information that they needed. There are many ways to provide shortcuts such as; history list, key shortcuts, and bookmarks.

*Services* aspect considers three sub-aspects: customizability, search function, and help services. These aspects refer to the support that assist and help the users during the navigation through e-learning system.

*Navigation aids* provide users helpful information and support during the navigation. Navigational structure, breadcrumbs, sitemaps and frames are few of the aids that provide constant feedback to users about where they are, and where they can go next in a particular web application.

*Consistency* is one of the important usability factors in any design endeavor and is equally important to enhance navigation speed in the context of LMS. Therefore, we take it as a separate aspect in our framework for good navigation structure. Inconsistency slows down the users to perform their tasks and find the relevant information. When users move through pages in any LMS, they should be provided with similar information, placement of menus/hyperlinks and accessibility of options in a similar way on every page. If consistency is kept in different modules in displaying the information it reduces navigation effort as users do not require wondering about how to move around.

The *adaptive navigation support* (ANS) is a set of techniques used to assist individual users to locate an appropriate navigation path to the goals and preferences in the context of hypermedia and hypertext, by adapting link presentation. ANS helps to overcome the limitations of static links presentations that provide the same set of links and page contents to all users, although users are relatively diverse by needs and preferences. Adaptive hiding and adaptive sorting are two sub factors in this category.

The primary goal of the proposed framework is to examine the efficiency of the navigation structure of any LMSs. By doing this, several navigational problems can be highlighted and improvements can be made to improve the navigation structure. This in turn leads to a better user experience and supportive learning environment.

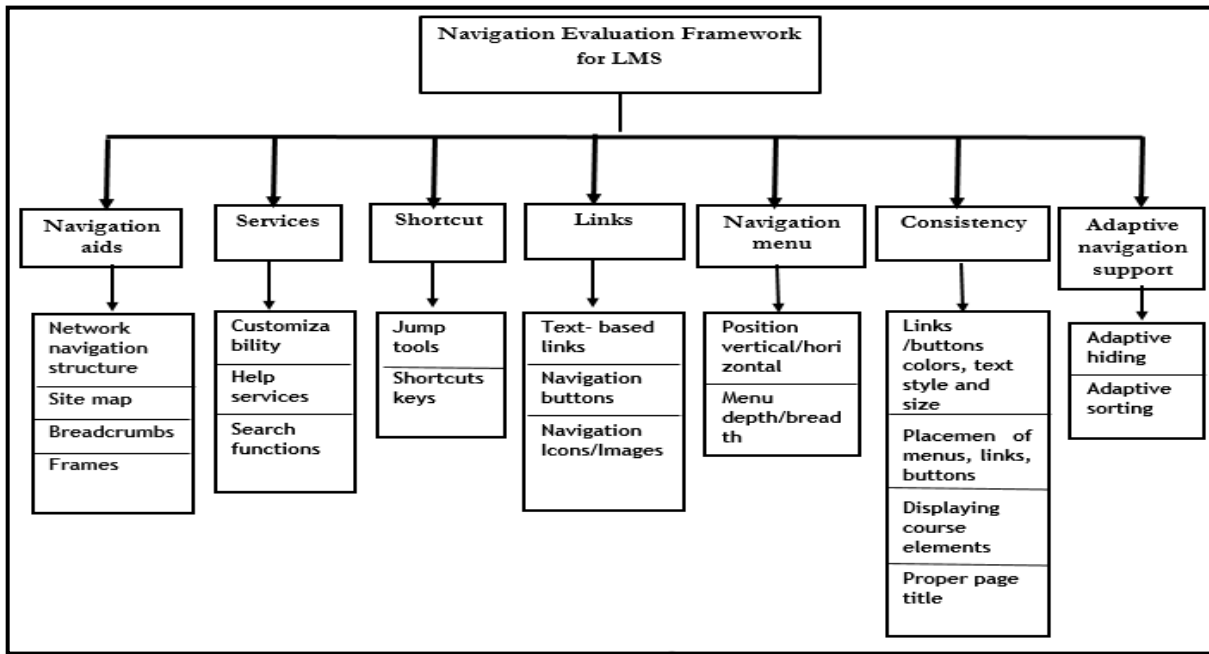


Fig. 1. Evaluation Framework

V. RESULTS AND ANALYSIS

Data collected through surveys is analysed using simple comparisons as well as advanced statistical techniques like Pearson correlation. This data was collected from two universities of Pakistan. For confidentiality reasons, these universities are referred as U1 and U2.

A. Analysis of generic questionnaire

The purpose of generic questionnaire was to get the opinion of students and teachers on the navigation structure of Moodle. Summary of the number of respondents on a generic questionnaire from both universities is presented in TABLE I.

Figure 2 shows a summary of usage of Moodle by students. It can be seen that most students of both universities use Moodle several times a week. Figure 3 indicates that most teachers of U1 used Moodle several times a week while at U2

TABLE I. NUMBER OF PARTICIPANTS IN GENERIC QUESTIONNAIRE

	U1			U2		
	Male	Female	TOTAL	Male	Female	TOTAL
Students	47	25	72	44	24	68
Teachers	2	2	4	4	2	6

Collective responses of Students = 140  
Collective responses of Teachers = 10

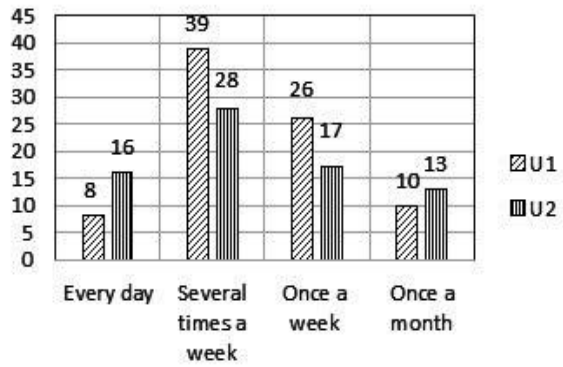


Fig. 2. Moodle usage by students

1) Links

TABLE II shows results of questions asked under the aspect links. Responses from both universities show that home link is available on every page of Moodle; no back to top link; icons and links/button text presented desired meaning to many of the respondents; no highlighted techniques are used; visited and unvisited links are not distinguishable.

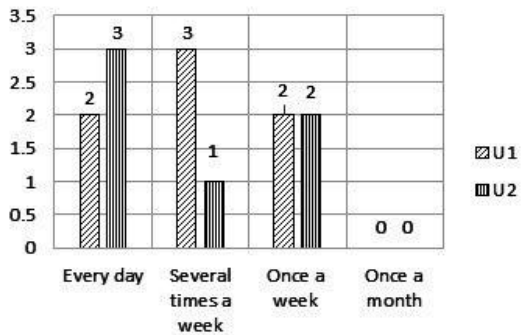


Fig. 3. Moodle usage by teachers

TABLE II. LINKS EVALUATION

Links	U1					U2					TOTAL					
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD	
Home Link	16	60	2	3	2	20	50	3	1	0	41	117	6	5	2	
	2	4	0	1	0	3	3	1	0	0						
Back to top link	10	15	3	40	16	15	13	6	18	22	26	29	10	64	41	
	0	1	1	3	2	1	0	0	3	1						
Icons and links/buttons text relevance	S	17	58	8	2	0	19	44	10	1	0	36	107	23	6	0
	T	0	0	4	3	0	0	5	1	0	0					
Use of highlighting techniques	S	6	30	36	9	2	12	21	27	12	2	19	58	66	22	5
	T	0	3	3	1	0	1	4	0	0	1					
Distinguishable visited and unvisited links	S	3	22	22	24	12	8	14	14	31	7	11	39	38	62	20
	T	0	1	1	4	1	0	2	1	3	0					

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree - S = Students, T = Teachers

2) Menu

TABLE III shows results of questions related to navigation menu. Results depict that most of the respondents found required items in navigation block and also the arrangement of items in the navigation block seemed to be logical, but still a few of them remained neutral that indicates menu items should be arranged in a more logical structure. Although many respondents reported that they were able to reach some particular destination with minimum clicks, but few of them reported that many steps are required to reach some sections.

TABLE III. NAVIGATION MENU EVALUATION

Navigation menu	U1					U2					TOTAL					
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD	
Availability of required option in navigation block	12	40	27	2	2	12	29	24	8	1	26	75	54	12	3	
	1	3	2	1	0	1	3	1	1	0						
Arrangements of navigation items is logical	S	5	52	17	7	2	12	35	19	6	2	17	96	39	14	4
	T	0	3	3	1	0	0	6	0	0	0					
Appropriate menu depth	S	4	47	21	5	6	15	26	16	13	4	19	81	42	18	10
		0	1	1	4	1	0	2	1	3	0					

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree - S = Students, T = Teachers

3) Navigational Aids

TABLE IV shows results of questions asked under the aspect navigational aids. The responses received on sub-aspects under this category reveal that mostly respondents were free to move in a non-sequential path; sitemap is not available in Moodle; bread crumbs are available; different sections of the screen are divided into different blocks.

TABLE IV. NAVIGATIONAL AIDS EVALUATION

Links	U1					U2					TOTAL					
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD	
Network navigation	11	48	14	8	2	12	33	18	10	1	24	85	38	20	3	
	0	3	4	0	0	1	1	2	2	0						
Sitemap	3	15	18	40	7	15	8	6	23	22	19	26	24	68	32	
	0	2	1	2	2	1	1	0	3	1						
Breadcrumbs	S	17	33	12	11	10	11	25	26	10	2	29	66	42	21	12
	T	1	4	2	0	0	0	4	2	0	0					
Frames	S	6	40	21	8	8	15	34	17	7	1	23	82	41	15	9
	T	1	6	0	0	0	1	2	3	0	0					

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree - S = Students, T = Teachers

4) Shortcuts facility

TABLE V shows results of questions asked under the aspect shortcuts facility. Results indicate that most students agreed that they are provided with the facility of the history list and bookmarks. However, most respondents remained neutral that shortcut keys are available in order to access different materials in Moodle.

TABLE V. SHORTCUTS FACILITY EVALUATION

Links	U1					U2					TOTAL					
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD	
Provision of history list	S	11	39	21	6	6	16	28	16	12	2	29	71	42	20	8
	T	1	2	3	1	0	1	2	2	1	0					
Facility of bookmarks	S	11	28	20	15	9	14	24	16	18	2	26	58	41	34	11
	T	0	3	4	0	0	1	3	1	1	0					
Shortcut keys facility	S	2	13	22	27	19	4	16	24	25	5	7	33	50	56	24
	T	0	1	1	4	1	0	2	1	3	0					

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree - S = Students, T = Teachers

5) Services

TABLE VI shows results of questions asked under the aspect shortcuts facility. Results indicate that most students remained neutral that they were able to customize page layout and courses according to their need and preference, while teachers agreed that they are provided with the service of customizability. Equal number of teachers agreed and remained neutral that help services give concise and easy-to-understand instructions.

Apart from teachers' responses, mostly students remained neutral that help services provide concise and easy to understand help instruction and also reported that there should be a direct help button. Most of the respondents strongly disagreed that they were able to find information, course, assignment, and lecture files, etc. through a search function, it means no powerful search function is available in Moodle.

TABLE VI. SERVICES EVALUATION

Links	U1					U2					TOTAL					
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD	
Customizability of page layout and most frequently visited courses	4	20	17	19	23	10	13	23	19	9	15	40	42	41	32	
	1	3	1	2	0	0	4	1	1	0						
Help services are concise and easy to understand	7	16	32	15	13	12	19	28	8	7	20	40	65	25	20	
	T	0	3	2	2	0	1	2	3	0		0				
Availability of search function	S	9	10	21	13	23	12	8	21	11	22	21	19	48	30	52
	T	0	1	3	3	0	0	0	3	3	0					

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree - S = Students, T = Teachers

6) Adaptive Navigation Support (ANS)

Pertaining to adaptive navigation, result (TABLE VII) depicts that most of the respondents disagreed that Moodle has support for adaptive navigation (adaptive sorting and adaptive hiding).

TABLE VII. ADAPTIVE NAVIGATION EVALUATION

Links	U1					U2					TOTAL				
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD
Adaptive hiding	2	15	15	21	30	9	5	10	24	26	12	26	27	49	56
	1	2	1	3	0	0	4	1	1	0					
Adaptive sorting	5	10	13	30	25	2	12	17	23	20	8	26	34	57	45
	T	0	2	2	3	0	1	2	2	1		0			

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree - S = Students, T = Teacher

7) Consistency

Results (TABLE VIII) under this aspect show that most respondents agreed that Moodle maintains consistency in links/buttons colors, text style and size. Some respondents remained neutral and that shows they did not find consistency in these elements throughout the use of Moodle. Again, most of the respondents found consistency in placement of menus, buttons and, links, but significant number of users remained neutral that shows they did not find consistency in placement of these navigation items. The result also depicts that most respondents agreed that there is consistency in displaying the course elements for different courses and every page of Moodle is properly titled. Despite this, some of the respondents reported that page title is not fixed in its location and moves up when the page is scrolled down.

TABLE VIII. CONSISTENCY EVALUATION

Links	U1					U2					TOTAL					
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD	
Consistency in colors, size, text style of navigation items	8	41	23	5	6	12	31	22	6	3	24	79	47	11	9	
	1	5	1	0	0	3	2	1	0	0						
Placement of navigation items is standard	12	26	37	4	4	11	40	20	3	0	23	74	59	10	4	
	0	4	1	2	0	0	4	1	1	0						
Placement of courses is consistent	S	11	46	18	4	4	14	37	20	1	2	26	91	41	6	6
	T	0	5	1	1	0	1	3	2	0	0					
Every page is titled properly	S	13	50	17	1	2	17	34	17	4	2	35	89	37	5	4
	T	3	2	2	0	0	2	3	1	0	0					

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree - S = Students, T = Teachers

8) Overall experience of navigation

Pertaining to the overall experience of navigation in Moodle we asked three questions in which respondents were required to rate their experience of ease of navigation, satisfaction about navigation design, and whether navigation design of Moodle helped accomplish tasks efficiently. Responses received against each question, depict that most respondents agreed that they experienced ease of navigation in Moodle easy. Results also show that most of the respondents were satisfied about navigation design of Moodle and agreed that navigation structure of Moodle helped them perform tasks efficiently (TABLE IX).

TABLE IX. OVERALL EXPERIENCE OF NAVIGATION

Experience of navigation	U1					U2					TOTAL				
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD
Experience of ease of navigation	8	42	22	8	3	16	30	23	3	2	26	79	49	11	5
	1	5	1	0	0	1	2	3	0	0					
Satisfaction about navigation design	7	35	22	16	3	11	33	21	6	3	20	75	45	24	6
	T	0	3	2	2	0	2	4	0	0		0			
Navigation helped perform tasks efficiently	4	38	29	10	2	16	31	21	5	1	20	77	54	16	3
	T	0	3	3	1	0	0	5	1	0		0			

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree - S = Students, T = Teachers

**B. Responses on subjective questions**

Subjective questions were also very important in order to assess the Moodle navigation structure in accordance with the student’s and teacher’s perspectives. This allowed the users to write any problems they faced and to provide suggestions for making the improvement in navigation design of Moodle.

First subjective question in our survey was “Describe any problems you encountered in navigation design of Moodle”. Users identified 30 problems and after closer review and classification, we found the following seven unique problems:

- 1) There is no back to top page button.
- 2) Sometimes users face difficulty in finding required content.
- 3) Navigation block has a lot of contents and font size is very small.
- 4) The information that has been displayed in a particular manner is confusing and increases stress on the mind.
- 5) Navigation is quite deep and sometimes requires a lot of steps to reach the required information/page.
- 6) Some links which are not relevant to users are still visible to them.
- 7) There is no direct Help button available.

The second subjective question in our survey was “Do you have any suggestions to improve the navigation structure of Moodle?” Some distinct answers from respondents were:

- 1) It should add video tutorials help for making easy for the students to use. For novice users’ it is quite difficult.
- 2) Links should be disable and enable according to the users’ requirements.
- 3) Links should be more structured and meaningful.
- 4) Icons should be more readable.
- 5) Include a helpful site map.
- 6) Navigational structure should highlight the important navigations.
- 7) A fixed menu on the left side would look and feel better.
- 8) Minimize the text and links in the navigation block.
- 9) Layout should change, it should add help feature.
- 10) Steps should be decreased for various sections.
- 11) Navigation structure should be improved by providing access to all pages from HOME menu, and buttons should be provided according to recently used history.

**C. Analysis of task based questionnaire**

In order to examine navigation of most frequently performed tasks in Moodle, this questionnaire was prepared. Each task had some set of questions, including not only the task to be performed, but also some questions regarding efficient navigation. The objective of this questionnaire was to get the users’ feedback on how they experienced Moodle in performing each task and also to reveal which tasks the participants were experiencing as easy or difficult. For students, we used lab-based usability testing method in which they were being observed while performing each task. Proper help was provided to students in case of any problem. Whereas, for teachers, we sent a questionnaire through e-mail

to them so they could fill it up on their workplace. Summary of the number of respondents from both universities is presented in TABLE X.

**1) Students’ perspectives – Usability test**

Students were required to perform following four tasks; upload and submit an assignment, access a lecture, access the weekly plan, and attempt a quiz in Moodle. The student sample included students from first year to fourth year, varying their level of experience with using the selected LMS. Figure 4 depicts the number of students participated in this usability test according to year of studies at both universities.

TABLE X. NUMBER OF PARTICIPANTS ON TASK- BASED QUESTIONNAIRE

	U1			U2		
	Male	Female	TOTAL	Male	Female	TOTAL
Students	47	25	72	44	24	68
Teachers	2	2	4	4	2	6
Collective responses of Students = 140						
Collective responses of Teachers = 10						

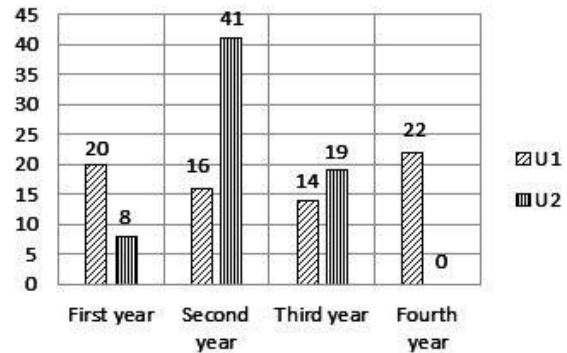


Fig. 4. Students’ educational experience (in terms of degree year)

Figure 5 represents the experience of students with the use of Moodle. This figure shows that most students of both universities were experienced with Moodle and results of overall response shows that most students had experience with the use of Moodle.

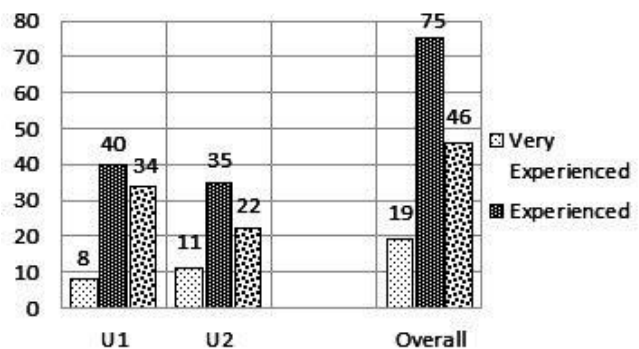


Fig. 5. Experience of students in using Moodle

**a) Task-one: Upload and submit an assignment**

**Result:** For Q1, the majority of students at both universities agreed that they could attach and submit an



assignment with minimum clicks. In Q2, the majority of students reported that button “Save Changes” presented desired meaning to them. But still some students reported that button text should be more meaningful as “Submit” or “Send”. For Q3, again majority of students from both universities agreed that submitted assignment status is shown in “RECENT ACTIVITY” block that reduces clicking effort, as they do not require to click particular assignment to check its status that whether it has been submitted for grading or not yet. For Q4, most students at both universities agreed that button/link text presented desired meaning to them (TABLE XI).

TABLE XI. RESPONSES ON TASK ONE

Q#	U1					U2					TOTAL				
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD
Q1	12	47	8	4	1	7	31	13	10	1	19	82	21	14	2
Q2	22	36	4	9	1	15	37	11	2	3	37	73	15	11	4
Q3	9	37	15	5	6	7	31	18	7	5	16	68	27	18	11
Q4	5	45	13	8	1	8	29	18	11	2	13	74	31	19	3

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree

b) Task-two: Download a Lecture

**Results:** For Q1 students at both universities reported that they were able to easily access recently uploaded lectures. Pertaining to Q2 majority of students from both universities agreed that files are easily downloaded. However, some students reported that sometime download links do not work properly. In Q3, we asked the students whether they mostly use “RECENT ACTIVITY” block to access uploaded lecture files and responses from students show that students at U1 were preferred this block while students at U2 disagreed with this, showing under course page they used to download lecture files. Lastly, for Q4, the majority of the students from both universities reported that minimum clicks were required to download the lecture files (TABLE XII).

TABLE XII. RESPONSES ON TASK TWO

Q#	U1					U2					TOTAL				
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD
Q1	18	40	9	4	1	13	39	9	5	2	31	79	18	9	3
Q2	20	39	11	1	1	21	29	12	3	3	41	68	23	4	4
Q3	7	23	20	13	9	9	16	10	22	11	16	39	30	35	20
Q4	17	42	12	1	0	21	30	7	7	3	38	72	19	8	3

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree

c) Task-three: View/Access weekly plan

**Result:** Pertaining to Q1, in which we asked whether they get notified when a new lecture, assignment or quiz is uploaded on Moodle, students of U1 remained neutral while students of U2 disagreed, the reason for these responses could be that students at U1 considered that upcoming events for particular course are shown in “Upcoming Events” block- a source to get users notified. Whereas, at U2, students disagreed because there was no such facility available from where students can check the upcoming events for their all

enrolled courses from a single page. Also, some of the students argued that to check upcoming events such as assignment, quizzes, and uploaded lectures they are required to check each course page separately- requiring many clicks. There should be a separate page that maintains a list of upcoming assignments, lectures, and quizzes for all subjects. Further, the result of this task indicates that majority of students agreed in Q2-Q3, and Q4 from both universities. This shows that students were required to perform minimum clicks to access the weekly plan, each week item is easily accessible under its respective course, and the current week is highlighted. In Q5 students at U1 agreed as they reported that in “Upcoming Events” block upcoming assignments and quizzes are shown with due dates and students at U2 were remained neutral in their responses but did not provide any supporting statements. Pertaining to Q6 most of the students disagreed from both universities indicated that weekly items, assignments and quizzes are not highlighted when submission date is near. Students stated that links should be highlighted when announcement for new lectures, assignment or quizzes is made; and when submission date is near for submitting the assignment and attempting the quiz. For Q7 students at U1 disagreed and students at U2 strongly disagreed that weeks’ dates with their items are able to expand and collapse (TABLE XIII).

TABLE XIII. RESPONSES ON TASK THREE

Q#	U1					U2					TOTAL				
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD
Q1	9	8	34	11	10	17	8	11	27	5	26	16	45	38	15
Q2	16	40	11	3	2	9	34	12	9	4	25	74	23	12	6
Q3	18	39	12	2	1	20	31	11	5	1	38	70	23	7	2
Q4	10	32	16	8	6	18	27	11	10	2	28	59	27	18	8
Q5	5	38	13	9	7	5	11	28	9	15	10	49	41	18	22
Q6	3	14	21	29	5	9	17	12	26	4	12	31	33	55	9
Q7	1	3	19	33	6	10	13	4	18	23	11	16	23	51	39

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree

d) Task-four: Attempt a quiz

**Result:** For Q1 students at both universities reported that they were able to easily access the quiz to be attempted. Pertaining to Q2 and Q3 majority of students from both universities agreed that they were able to navigate through different questions of the quiz. For Q4, the majority of students from both universities reported that they were able to finish their quiz on any question. For Q5 again majority of students from both universities reported that location of buttons/links in this module is easily noticeable. For Q6 at U1 most students agreed that button/link text presented desired meaning to them. While at U2, in response to this question majority of students remained neutral (TABLE XIV).

TABLE XIV. RESPONSES ON TASK FOUR

Q#	U1					U2					TOTAL				
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD
Q1	20	36	12	1	3	16	31	13	6	2	36	67	25	7	5

Q2	13 32 11 9 7	16 27 14 8 3	29 59 25 17 10
Q3	9 31 18 8 6	14 24 22 6 2	23 55 40 14 8
Q4	10 38 10 9 5	13 24 18 10 3	23 62 28 19 8
Q5	16 38 6 10 2	11 28 17 8 4	27 66 23 18 6
Q6	6 26 17 19 4	7 19 23 14 5	13 45 40 33 9

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree

After accomplishing all the tasks students were asked to rate the level of difficulty in accomplishing each task, and results indicate that most of the students of both universities found easy to perform all tasks in Moodle (TABLE XV).

TABLE XV. LEVEL OF DIFFICULTY IN PERFORMING EACH TASK

Q#	U1					U2					TOTAL				
	VE	E	N	D	VD	VE	E	N	D	VD	VE	E	N	D	VD
Q1	12	42	16	2	0	10	33	21	1	3	22	75	37	3	3
Q2	21	35	15	1	0	17	29	15	4	3	38	64	30	5	3
Q3	14	31	18	7	2	13	30	16	6	3	27	61	34	13	5
Q4	17	34	20	1	0	15	30	19	3	1	32	64	39	4	1

VE = very easy, E = easy, N = neutral, D = difficult, VD = very difficult

**Observed problem:** The problem observed was that mostly first year students who had no prior or little experience of using Moodle, did not able to know how to attach an assignment for assessment and there was no proper help available in Moodle. The majority of the students struggled with understanding the Moodle structure of that page, and they kept asking about how to upload a file.

2) Teachers' perspectives

Teachers were required to perform the following tasks, upload assignment, upload a lecture, modify weekly plan, and make a quiz in Moodle while each had further sub-tasks. After accomplishing each task they were required to respond on providing a task based questionnaire. Only 4 teachers from U1 and 6 from U2 responded back the questionnaire – a total of 10 responses were received from teachers (Table 4.48). Nielsen [30] stated that testing five users is typically enough for evaluating design, most importantly the usability problems. Hence, we considered this size enough to find out teachers' problems about navigation design of Moodle. Figure 6 represents the experience of teachers with the use of Moodle.

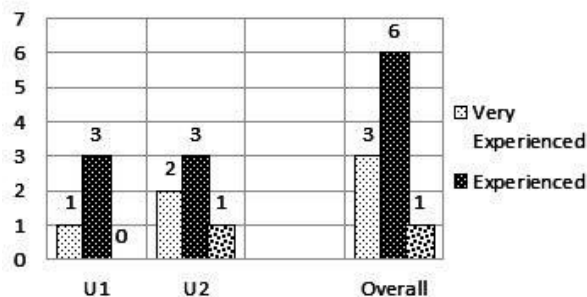


Fig. 6. Experience of using Moodle (Teachers)

a) Task One: Upload an assignment

**Results:** In response of Q1 teachers at U1 responded as neutral, but teachers at U2 agreed that they needed help to upload an assignment. In Q2 it was easy for teachers of U2 to set assignment submission type as "Online text", however, teachers of U1 teachers remained neutral. For Q3, description of links in performing this task presented desired meaning to teachers at both universities. Pertaining to Q4, teachers found that minimum steps were required to upload an assignment (TABLE XVI).

TABLE XVI. RESPONSES ON TASK ONE

Q#	U1					U2					TOTAL				
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD
Q1	1	2	1	0	0	1	3	0	1	1	2	5	1	1	1
Q2	1	0	2	1	0	0	4	0	1	1	1	4	2	2	1
Q3	1	2	1	0	0	1	3	2	0	0	2	5	3	0	0
Q4	0	2	2	0	0	1	3	2	0	0	1	5	4	0	0

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree

b) Task Two: Upload a Lecture

**Results:** For Q1, teachers reported that they were able to attach multiple files at a time. For Q2, teachers at U1 agreed that label of buttons showed the desired destination. In case of teachers of U2, equal number of teachers were agreed and neutral. For Q3, teachers of U1 easily found how to hide a file, while at U2, equal number of teachers agreed and disagreed that they found the option for hiding the file easily. While for Q4, teachers at both universities reported that they easily renamed the file. Lastly, for Q5, links/selection presented desired meaning to teachers of both universities (TABLE XVII).

TABLE XVII. RESPONSES ON TASK TWO

Q#	U1					U2					TOTAL				
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD
Q1	0	1	1	2	0	0	4	1	1	0	0	5	2	3	0
Q2	0	2	2	0	0	1	4	0	1	0	1	6	2	1	0
Q3	0	1	1	2	0	0	4	1	1	0	0	5	2	3	0
Q4	0	2	2	0	0	0	3	2	1	0	0	5	4	1	0
Q5	1	2	1	0	0	2	2	1	1	0	3	4	2	1	0

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree

c) Task Three: Modify weekly plan

**Results:** Result of this task shows that teachers at both universities easily found delete option as a response to Q1. For Q2, 2 teachers of U1 disagreed that links/ icons presented desired meaning to them, and teachers of U2 remained neutral. Pertaining to Q3, most teachers of both universities reported that they were not able to modify weekly plans and needed help. For Q4, 2 out of 4 teachers at U1 remained neutral, while at U2, teachers found it easy to hide a week. For Q5, most teachers at both universities agreed that the current week is highlighted. Pertaining to Q6, teachers of both universities

responded that they were unable to expand and collapse weekly plan. For Q7 teachers of both universities remained neutral and reported that very often they remember that they need to click “Turn editing on” button in order to add/modify course elements (TABLE XVIII).

TABLE XVIII. RESPONSES ON TASK THREE

Q #	U1					U2					TOTAL				
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD
Q1	0	3	1	0	0	1	3	1	1	0	1	6	2	1	0
Q2	0	1	1	2	0	0	1	3	2	0	0	2	4	4	0
Q3	0	2	1	1	0	0	4	1	1	0	0	6	2	2	0
Q4	0	1	2	1	0	1	3	2	0	0	1	4	4	1	0
Q5	0	3	0	1	0	0	4	2	0	0	0	7	2	1	0
Q6	0	0	1	3	0	0	0	2	4	0	0	0	3	7	0
Q7	0	0	3	1	0	0	1	3	2	0	0	1	6	3	0

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree

d) Task Four: Make a quiz

**Results:** Responses on Q1 were as teachers of U1 disagreed that they feel it simple to make a quiz in Moodle while at U2, teachers reported that it is easy to make a quiz in Moodle. For Q2, at U1, equal number of responses received on agreed and neutral that link/button’s text helped in recognizing how to add quiz in Moodle. Whereas, at U2, teachers agreed with this question. For Q3, teachers of U1 disagreed that there are certain confusing link descriptions. While teachers of U2 faced some confusing link descriptions. For Q4, equal number of teachers at U1 agreed and remained neutral, whereas at U2, teachers agreed that there are many steps to make a quiz which can be eliminated. For last question that is Q5, teachers of U1 reported that they needed help to add a quiz in Moodle while at U2 teachers agreed that they required help to add a quiz in Moodle (TABLE XIX).

TABLE XIX. RESPONSES ON TASK FOUR

Q#	U1					U2					TOTAL				
	SA	A	N	D	SD	SA	A	N	D	SD	SA	A	N	D	SD
Q1	0	0	1	3	0	1	1	0	3	1	1	1	1	6	1
Q2	0	3	1	0	0	3	3	0	0	0	3	6	1	0	0
Q3	0	2	2	0	0	0	3	0	3	0	0	5	2	3	0
Q4	1	2	1	0	0	1	4	1	0	0	2	6	2	0	0
Q5	0	3	0	1	0	0	5	1	0	0	0	8	1	1	0

SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree

After accomplishing all the tasks, teachers were asked to rate the level of difficulty in accomplishing each task, and the results indicate that the majority of teachers reported that they found it easy to perform tasks one, two, and three. However, task four which was related to the creation of a quiz was difficult and required help.

TABLE XX. LEVEL OF DIFFICULTY IN PERFORMING EACH TASK

Q#	U1					U2					TOTAL				
	VE	E	N	D	VD	VE	E	N	D	VD	VE	E	N	D	VD
Q1	12	42	16	2	0	10	33	21	1	3	22	75	37	3	3
Q2	21	35	15	1	0	17	29	15	4	3	38	64	30	5	3
Q3	14	31	18	7	2	13	30	16	6	3	27	61	34	13	5
Q4	17	34	20	1	0	15	30	19	3	1	32	64	39	4	1

VE = very easy, E = easy, N = neutral, D = difficult, VD = very difficult

D. Results of correlation

To answer the hypotheses H1 and H2, we performed Pearson correlation analysis. For correlation tests responses of both universities were collectively analyzed. The purpose of H1 is to find out whether ease of navigation correlates with satisfaction about navigation design and efficiency of performing task Results (TABLE XXI) indicate that ease of navigation has a significant relationship with satisfaction about navigation design i.e. 0.803 and efficiency of completing the task i.e. 0.989. It means, if a user perceives ease of navigation, it increases the user’s satisfaction about navigation design.

TABLE XXI. CORRELATION OF EASE OF NAVIGATION WITH SATISFACTION ABOUT NAVIGATION DESIGN AND EFFICIENCY OF PERFORMING THE TASKS

Factors	Satisfaction about Navigation Design	Efficiency of Performing Task
Ease of Navigation	0.803	0.989

TABLE XXII presents the result of H2. The purpose of H2 is to check whether seven factors of the framework are correlated to ease of navigation experience and satisfaction about navigation design. Result of correlation confirms the positive relationship of all navigation factors with ease of navigation and satisfaction about navigation design. The result shows that links are significantly correlated with ease of navigation and satisfaction about navigation design. It means that if link appearance indicates that it is clickable and text of the link is related to visitors of the page, it will have a positive effect on ease of navigation and satisfaction about navigation design. The menu is also positively correlated with ease of navigation and satisfaction about navigation design having values 0.719 and 0.862 respectively, indicating that if users perceive that items in a menu are arranged logically, it enhances their ease of navigation experience and satisfaction on navigation design. Moreover, result in table shows that all other variables such as navigation aids, shortcuts facility, services facility, adaptive navigation support, and consistency are also strongly correlated with ease of navigation and satisfaction about navigation design.

TABLE XXII. CORRELATION OF NAVIGATION FACTORS WITH EASE OF NAVIGATION AND SATISFACTION ABOUT NAVIGATION DESIGN

Factors	Ease of Navigation	Satisfaction about Navigation Design
Links	0.984	0.801
Menus	0.719	0.862

Navigation Aids	0.867	0.826
Shortcuts	0.967	0.726
Services	0.870	0.820
Adaptive Navigation Support	0.785	0.886
Consistency	0.839	0.861

## VI. CONCLUSIONS AND FUTURE WORK

Moodle is selected for evaluation in order to explore whether it contains all navigational factors of proposed navigational evaluation framework for LMS or how well these are maintained in Moodle. Surveys result revealed that Moodle does not include some navigational factors and those are implemented requiring improvements. The navigation factors of proposed framework that are not provided in Moodle are back to top link, site map, proper help and documentation, powerful search function, and shortcut key facility. In addition, there are no highlighting techniques implemented in Moodle, which can direct the users to important information or events. These are all essentials for an efficient navigation. Furthermore, customizability is also an important aspect of good navigation, but Moodle still does not allow users to customize pages/courses according to their need and preference. Some of the users are not satisfied that links are still visible to them even though these are not relevant to them.

Moreover, some of the problems encountered in the navigation design of Moodle are inconsistency in color and placement of buttons on different pages, presence of back page button on some pages but on some pages there is no back page button, there is no strategy by which users can differentiate visited and not-visited links, some of the icons are not understandable just by looking at them and there is no supporting text that help understanding the desired meaning of icons, unnecessary options are available in navigation block (menu), menu items change when moving to a new page, for some users organization of items in navigation block does not seem logical that is rarely required items are placed at the top.

Navigation of LMSs has great influence on learning and teaching processes. If navigation of such systems is not efficient it can distract students from actual learning because they will spend most of the time in learning how to use the system. Also, if LMSs hinder navigation, teachers will be spending most of the time in trying to understand the system and giving less time for teaching activities, since they would have already been frustrating in understanding the way systems works. Hence, in order to overcome these obstacles a proper navigation structure should be maintained to enhance learning and teaching activities. By doing this, we are making the users' satisfied and increasing their experience of using LMS.

Furthermore, results of correlation showed that ease of navigation is positively correlated with the factors of proposed framework. Also, correlation result indicated that ease of navigation is positively correlated with efficiency of achieving

goals. It means users feel ease of navigation and it increases their efficiency of achieving the goal quickly.

In this research, we had insufficient sample size of teachers to reach some reliable results in perspective of teachers. Moreover, our sample set only included students from strong IT background. In future, we aim to increase target sample size in which respondents from business, law, environmental sciences and from some other disciplines will be included. In addition, sample size of teachers will be increased in future research. In future, further, we plan to investigate: Moodle's navigation structure in comparison to other open source LMSs based on proposed framework; and how different users' groups based on gender, age, educational background, and computer proficiency, perceive navigation in LMS.

## REFERENCES

- [1] S. Kim and M. Lee, "Validation of an evaluation model for learning management systems," *J. Comput. Assist. Learn.*, vol. 24, no. 4, pp. 284–294, 2008.
- [2] S. Minocha and H. Sharp, "Learner-Centred Design and Evaluation of Web-Based E-Learning Environments" in *The 7th HCI Educators Workshop: Effective Teaching and Training in HCI*, University of Central Lancashire, Preston. 1-2 April, 2004.
- [3] Y. Hammouche, "The Usability Testing of Optima Learning Management System – From Student Interface," *Turku University of Applied Sciences*, 2014.
- [4] P. Zaharias, "A usability evaluation method for e-learning," in *CHI '06 extended abstracts on Human factors in computing systems - CHI '06*, 2006, no. January 2006, p. 1571.
- [5] P. Zaharias and A. Poylymenakou, "Developing a Usability Evaluation Method for e-Learning Applications: Beyond Functional Usability," *Int. J. Hum. Comput. Interact.*, vol. 25, no. 1, pp. 75–98, 2009.
- [6] A. Shahzad, W. Hafizi, and A. G. Golamdin, "E-Learning Navigation Model Based on Student's Learning Behavior : Case Study in UUM," *Int. J. Bus. Soc. Sci.*, vol. 5, no. 5, pp. 118–130, 2014.
- [7] A. M. Reina, "Separating the Navigational Aspect," in *Proc. 2nd Int'l Workshop on Aspect Oriented Programming for Distributed Computing Systems (ICDCS-2002)*, Vol. 2, 2002.
- [8] E. Zakaria and M. Y. Daud, "The Role of Technology: Moodle as a Teaching Tool in a Graduate Mathematics Education Course," *Asian J. Manag. Sci. Educ.*, vol. 2, no. 4, pp. 46–52, 2013.
- [9] J. Nielsen, "Usability 101: Introduction to Usability," Nielsen Norman Group, 2012. [Online]. Available: <http://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/>. [Accessed: 04-Aug-2015].
- [10] "Usability", Wikipedia. [Online]. Available: <https://en.wikipedia.org/wiki/Usability>. [Accessed: 02-Jun-2015].
- [11] A. Inversini, L. Botturi, and L. Triacca, "Evaluating LMS Usability for Enhanced eLearning Experience," *Ed-Media 2006*, pp. 595–601, 2006.
- [12] Y. Wang, H. Wang, W. Huang, and Y. Jiao, "Implementing a web-based personalized learning and assessment system," *ICCSE 2010 - 5th Int. Conf. Comput. Sci. Educ. Final Progr. B. Abstr.*, pp. 1629–1632, 2010.
- [13] J. Reyna, "Developing quality e-learning sites : A designer approach," *New York*, pp. 837–838, 2009.
- [14] "Why you must free up navigation in e-learning." [Online]. Available: <http://info.shiftelearning.com/blog/bid/241958/3-Reasons-Why-You-Must-Free-Up-Navigation-In-eLearning>. [Accessed: 6-Jun-2015].
- [15] C. McKnight, J. Richardson, and A. Dillon, "Hypertext: a psychological perspective," *Ellis Horwood Series in Interactive Information Systems Published by Ellis Horwood*. 1993.
- [16] J. Melton, "The lms moodle: A usability evaluation," *Prefect. Univ. Kumamoto Retrieved ...*, pp. 1–24, 2006.

- [17] G. Kakasevski, "Evaluating usability in learning management system Moodle." *Information Technology Interfaces, 30th International Conference on. IEEE*, 2008.
- [18] M. Ivanović, Z. Putnik, Ž. Komlenov, T. Welzer, M. Hölbl, and T. Schweighofer, "Usability and privacy aspects of moodle: Students' and teachers' perspective," *Inform.*, vol. 37, no. 3, pp. 221–230, 2013.
- [19] M. A. Amaral and C. M. De Almeida, "Usability Assessment in Moodle Abstract : 2 E-learning and Usability," vol. 1, no. 10, pp. 1–10, 2009.
- [20] "LMS Usability Evaluation Report," Prepared by the LMS Usability Working Group, 2010.
- [21] I. Padayachee, P. Kotzé, and A. van der Merwe, "Course Management Systems from a Usability Perspective," *Alternation*, vol. 18, no. 1, pp. 297 – 317, 2011.
- [22] L. Martin, D. R. Martínez, O. Revilla, M. José, O. C. Santos, and J. G. Boticario, "Usability in e-Learning Platforms : heuristics comparison between Moodle , Sakai and dotLRN," *Artif. Intell.*, vol. 509, no. L11, pp. 75–84, 2003.
- [23] D. Zub and E. Eessaar, "Technological Developments in Education and Automation," 2010.
- [24] J. Rosato, C. Dodds, and S. Laughlin, "Usability of Course Management Systems by Students," *Methods*.
- [25] M. Machado and E. Tao, "Blackboard vs. Moodle: Comparing user experience of learning management systems," *Proc. - Front. Educ. Conf. FIE*, no. December 2006, pp. 7–12, 2007.
- [26] H. Baytiyeh, "Work in progress; Open source usability evaluation: The case of Moodle," *2011 Front. Educ. Conf.*, pp. F4E–1–F4E–2, 2011.
- [27] W. Gate and R. Heights, "A usability study of moodle," no. 2007, pp. 60–84, 2013.
- [28] A. Seffah, T. Mohamed, H. Habieb-mammar, and A. Abran, "The Journal of Systems and Software Reconciling usability and interactive system architecture using patterns," vol. 81, pp. 1845–1852, 2008.
- [29] A. Oztekin, Z. James, and O. Uysal, "International Journal of Industrial Ergonomics UseLearn : A novel checklist and usability evaluation method for eLearning systems by criticality metric analysis," *Int. J. Ind. Ergon.*, vol. 40, no. 4, pp. 455–469, 2010.
- [30] "Usability 101: Introduction to Usability." [Online]. Available: <http://www.nngroup.com/articles/usability-101-introduction-to-usability/>. [Accessed: 1-Oct-2015].

# Communication-Load Impact on the Performance of Processor Allocation Strategies in 2-D Mesh Multicomputer Systems

Zaid Mustafa,

Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), University of Santiago de Compostela, Spain

J. J. Alshaer

Department of Computer Information Systems,  
Albalqa Applied University  
Alsalt, Jordan

O. Dorgham

Department of Computer Information Systems,  
Albalqa Applied University  
Alsalt, Jordan

S. Bani-Ahmad

Department of Computer Information Systems,  
Albalqa Applied University  
Alsalt, Jordan

**Abstract**—A number of processor allocation strategies have been proposed in literature. A key performance factor that can highlight the difference between these strategies is the amount of communication conducted between the parallel jobs to be allocated. This paper aims to identifying how the density and pattern of communication can affect the performance of these strategies. Compared to the work already presented in literature, we examined a wider range of communication patterns. Other works consider only two types of communication patterns; those are the one-to-all and all-to-all patterns. This paper used the C language to implement the different allocation strategies, and then combined them with the ProcSimity simulation tool. The processor-allocation strategies are examined under the First-Come-First-Serve scheduling strategy. Results show that communication pattern and load are effective factors that have a significant impact on the performance of the processor allocation strategy used.

**Keywords**—Processor allocation; Parallel computing; 2-D Mesh; Communication patterns; Multicomputer systems

## I. INTRODUCTION

Nowadays, Multi-computer computer systems are widely used, because they are cost-effective alternatives of the traditional supercomputers [8]. The Topology of a multi-computer defined as a style based on the processing units of interconnected multi-computer. Mesh-based topologies considered as one of these topologies, such as the two-dimensional (2-D) and the three-dimensional (3-D) topologies. These topologies are the most common topologies; according to their simple, regular and scalable [1, 7,4].

Most of the modern commercial and experimental multi-computer systems (such as are the IBM BlueGene/L and the Intel Paragon) have been built based on this architecture [8, 4].

In a multi-computer system, the processor allocator applies processor allocation strategies in order to assign unallocated multi-computer nodes to parallel jobs and identify them [8]. Both of processor allocation algorithms and better idle-nodes

recognition ability for multi-computers or idle sub-meshes can reduce the job waiting time and improve the chance of assigning a parallel job into the system in a significant way [8]. The proper choice of the processor allocation strategy in 2-D-Mesh multi-computer consider as a critical issue in deciding the performance of a given multi-computer system. As it is significantly affects the performance (system utilization, response time, throughput ... etc) of any parallel system [8, 4].

Allocation strategies fall into two main categories, as follows:

- First: contiguous allocation strategies seek to allocate a sub-mesh, such as a contiguous set of processing units of the same size and shape of parallel request. Contiguity condition can be summarized in the relaxed of the non-contiguous allocation strategy [4]. Contiguous allocation strategies have low system utilization, high internal and external fragmentation, which consider as a problem to the contiguous allocation strategies.
- Second: non-contiguous strategies which can produce high execution times of parallel jobs according to high communication latencies [4, 5, 6].

Fragmentation problem prevents an idle processors from being utilized. Which can be classified into internal and external fragmentation. Internal fragmentation is the result of allocating jobs only to certain size sub-meshes. It occurs when a job is assigned to more processors than it requires, or extra nodes allocated are not used for actual computation [8]. Also, internal fragmentation occurs when a job requests a sub-mesh that does not fit the requirement of the allocation algorithm. For example, internal fragmentation occurs when a job does not require a square sub-mesh with sides equal to power of two and is allocated using the TDB scheme. The external fragmentation happens when the allocation scheme cannot allocate the available processors to the incoming jobs [8].



In addition to, the non-contiguous allocation strategies tries to fix the problems of external and internal fragmentation, beside the low system utilization by allowing parallel requests to be partitioned and allocated non-contiguously into smaller sub frames in case contiguous allocation fails [8, 4, 3]. Studies showed that allocation strategies integrated the advantages of both contiguous and non-contiguous allocation strategies throughout preserving some level of contiguity within allocated parallel job [3].

In this paper, the authors comparatively evaluates a number of partially non-contiguous allocation strategies that are proposed in [4, 3] in terms of their performance at high communication loads. A number of communication styles will be experimented. The allocation strategies of interest will be implemented in the C language, and later integrated with the ProcSimity simulation tool [12, 16]. The processor allocation strategies will be examined under the scheduling strategy called First-Come-First-Serve as in [7, 3].

## II. PROBLEM STATEMENT AND EXPERIMENTED COMMUNICATION PATTERNS

In this paper, contiguous and partially non-contiguous allocation strategies will be comparatively evaluated in terms of their performance at high communication loads (a number of communication styles will be experimented [4, 3]). The goal is to compare the allocation strategies in terms of their performance at *high communication loads*.

This paper tries to answer the following questions:

- Does the behavior of the processor allocation strategy in use change as the communication behavior of parallel jobs change?
- Does the behavior of the processor allocation strategy used change as the communication load caused by the parallel jobs change?

One communication scheme to be tested is the Fast Fourier Transform (FFT) pattern. The FFT algorithm is introduced as an efficient algorithm to compute the discrete Fourier transform "DFT" and it is inverse [19].

Another important communication patterns is the NAS parallel Bench marks (NPB) which represents a small set of programs designed to help evaluating the performance of parallel supercomputers. The benchmarks are derived from computational fluid dynamics (CFD) applications and consist of five Kernals and three pseudo-applications. The original benchmarks identified in NPB mimic the computation and data movement in CFD applications, the five Kernals is IS, EP, CG, MG,FT; and the three pseudo applications is BT, SP, LU [20; 21; 23].

The authors also tested the Divide-and-Conquer communication pattern. A Divide-and-Conquer-based algorithm solves a programming problem by recursively partition that problem into sub-problems of roughly equal size, if sub-problem can be solved independently; there is a possibility increase the speed up by parallel computing [24].

## III. LITERATURE REVIEW AND RELATED WORK

The performance of parallel systems can be significantly affected by the processor allocation strategy used in the 2D-Mesh multicomputer system in hand [1]. A number of studies have proposed allocation strategies. All proposed strategies can fall under two categories: namely; contiguous and non-contiguous strategies.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

### A. Contiguous Processor Allocation Strategies

In contiguous processor allocation of a given parallel request, a contiguous set of processing units of the same size and shape of that request are assigned to job of interest [17, 1].

In contiguous allocation strategies a low level of system utilization can be raised and external and/or internal fragmentation can be observed [8,11]. When a job needs to be allocated to a number processor, it may be assigned more processing units than what it actually requires. This is referred to as the internal fragmentation problem [8,7]. External fragmentation on the other hand, occurs when enough number of idle processors is available in the system but cannot be allocated to the scheduled parallel job because of the necessity of contiguity [1,4]. Many research efforts tried to solve or decrease the problem of external fragmentation [8,4], and one proposed solution was to use non-contiguous allocation strategies.

All the above allocation strategies are referred to as contiguous allocations because they consider only contiguous regions for the execution of a parallel job. In contiguous allocation, communication cost is minimal [6]. However, the requirement that a parallel job has to be allocated to contiguous set multi-computers reduces the chance of successfully allocating the job due to the problem of fragmentation [6,7].

Two Dimensional Buddy (2DB) [25]: in the Two Dimensional Buddy (2DB) strategy, the system is assumed to be a square with side lengths equal to a power of two ( $2^i$ , where  $i=0, 1, 2, \dots$ ). The size of a requested sub-mesh is rounded up to a square with side-lengths as the nearest power of two. The resulting sub-mesh after rounding up can be larger than the original sub-mesh. Consequently, this allocation strategy suffers from internal fragmentation because it only allocates square sub-meshes whose side lengths are equal to a power of two. In other words, because of rounding up the sides of requests, the allocated sub-mesh can be larger than the originally requested sub-mesh.

Frame Sliding (FS) [8]: The Frame Sliding (FS) method was proposed to reduce the internal fragmentation problem of the 2DB allocation. This is achieved through allowing sub-meshes of any arbitrary size to be allocated to parallel jobs. This allocation strategy works as follows: Viewing the requested sub-mesh of the job in hand as a frame, the FS algorithm slides the frame across the system to examine for a free sub-mesh to execute the job [8].

The First-Fit (FF) and the Best-Fit (BF) Strategies [18]: The FF and BF algorithms guarantee the recognition of a free sub-mesh, provided it exists. The two algorithms work by scanning the entire mesh for possible allocation.

Adaptive-Scan (AS) Strategy [9]: The adaptive-scan changes the orientation of the sub-mesh being searched for if the required sub-mesh in the original orientation is not available. Thus, the AS strategy has better recognition capabilities than that of the BF and FF schemes.

#### B. Non-contiguous Processor Allocation Strategies Processor

When using non-contiguous allocation the contiguity condition is not a must [8]; therefore, a job does not have to wait for a single sub-mesh of the requested size and shape to be available because it will be executed on a number of disjoint smaller sub-meshes [8; 15].

The non-contiguous allocation of requests can successfully solve the drawbacks of contiguous allocation strategies. As experimentally validated, non-contiguous allocation strategies have been found to produce relatively high system utilization and eliminate fragmentation [8, 15, 14].

Communication latency is often high in non-contiguous allocation strategies since communication between processors running the same job may be indirect due to non-contiguity [14].

The introduction of wormhole routing [10] encouraged researchers to consider non-contiguous allocation on multicomputers [8]. With wormhole routing, the message communication latency is less dependent on the distance traveled by the message from source to destination [8, 10].

In non-contiguous allocation schemes, allocation requests of parallel jobs are subdivided into two partitions [8] if contiguous allocation fails. If allocating either of the two partitions fails, that partition is further subdivided into smaller sub-partitions until the allocation succeeds [11]. Using wormhole routing has made allocating parallel jobs to non-contiguous processors reasonable in terms of performance even in networks characterized by a relatively long-diameter, "The diameter is the shortest hop "the maximum of the shortest distance between any two nodes", a small communication delay happens when an interconnection graph has a small diameter between nodes [22]. Such the 2-D mesh. So the contiguity condition is relaxed which allowed jobs to be executed without waiting for sufficient and contiguous set of idle processing nodes to be available in the system [7; 3; 11].

When "Wormhole routing" is a special case of cut-through switching. It routes the head of a packet directly from incoming to outgoing channels of the routing chip. A packet is divided into a number of flits (i.e. flow control digits) for transmission. The route or the path followed by those flits from source to destination is determined by the header flits (or flits). Flits are forwarded through a chosen channel after examining the header flit(s) of a message. As the header flit of a given packet/message advances along a specific route, all following flits follow in through the same route. [13].

There are many important strategies of the non-contiguous allocation such as Multiple Buddy System (MBS) expresses the allocation request as a base-4 number, and bases allocation on this expression. In this strategy, the mesh of the system at hand is divided into non-overlapping square sub-meshes with side lengths that are powers of 2. The number of processors,  $p$ , requested by a scheduled job is factorized into a base-4 block. If a required block is unavailable, MBS recursively searches for a larger block and repeatedly breaks it down into four buddies until it produces blocks of the desired size. If that fails, the requested block is further broken into four sub-requests until the job is allocated [3].

In the Paging allocation strategy, for instance [2], the entire 2D mesh is virtually sub-divided into pages or sub-meshes of equal sides' length of  $2i$  where  $i$  is a positive integer Number that represents the index parameter of the paging Approach. That can scan for pages in multiple ways such as that snake-line order and row-major order [3].

A main disadvantage of this strategy is the fragmentation it may cause when some free processors' cannot be allocated because they are contained in pages that have been reserved to job [13].

A partially non-contiguous allocation strategy tries to find a contiguous set of processing units of the same size to active job using some contiguous strategy. If it fails, the active job is divided into two sub-requests. The two new sub-requests are then allocated using the same contiguous allocation approach again; this operation will continue recursively until the request is finished. Examples of this allocation strategies are (i) the PALD-FF (PARTitioning at the Longest Dimension with First-Fit), (ii) the PALD-BF (PARTitioning at the Longest Dimension with Best-Fit) [3] and (iii) the restricted size reduction (RSR) strategy [3]. The RSR strategy allows jobs to be executed on a reduced size sub-mesh adaptively and partitions it in two sub-blocks of equal sizes.

In the PALD\_FF strategy, First-Fit approach used to find a contiguous group of processing units of the same size and shape of the application at hand. In case of fail, the request at hand is divided into sub requests after removing one from the longest dimension of the request that is for a given request of size  $a*b$  and assuming  $b>a$ , the two partition-sizes are  $a*(b-1)$  and  $a*1$  after removing one from the longest dimension of the request. The two new sub-requests are then allocated using the First-Fit approach again; this procedure continues recursively until the request is fulfilled [3].

According to the previous, the PALD\_BF is the same of PALD\_FF but it applies the best fit strategy rather than the first fit. processor allocation of a given parallel request, a contiguous set of processing units of the same size and shape of that request are assigned to job of interest [17, 1].

#### IV. RESULTS AND OBSERVATIONS

Figures 1 through 5 show the relationship between the system load from one side, and, from the other side, (i) the mean job response time, (ii) the mean job service time, (iii) the mean packet blocking time, (iv) the mean packet latency, and (v) the percent system utilization.

Figures 1 through 5 are specific for the case of having all-to-all communication pattern. The allocation strategies used in these figures are: (i) the Random, (ii) the Multiple Buddy System, (iii) the First-Fit, and (iv) the Best-Fit allocation strategies. The first two strategies are non-contiguous strategies, while the last two are contiguous. The following measures are used to evaluate allocation techniques: MJRT (Mean Job Response Time), MJST (Mean Job Service Time), MPBT (Mean Packet Blocking Time), MPL (Mean Packet Latency) and PSU (Percent System Utilization). The following Figures show results of simulation. Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable. Figure 1 represents the relationship between the system load and mean job response time with the allocation strategies "MBS, FF, BF, Random" under the first come first serve scheduling mechanism and all-to-all communication pattern. The communication load parameter here is the mean number of messages sent by any scheduled job. This parameter has made fixed at 80. Observation 1 (figure 1): In general, contiguous allocation strategies show higher MJRT than the non-contiguous allocation strategies.

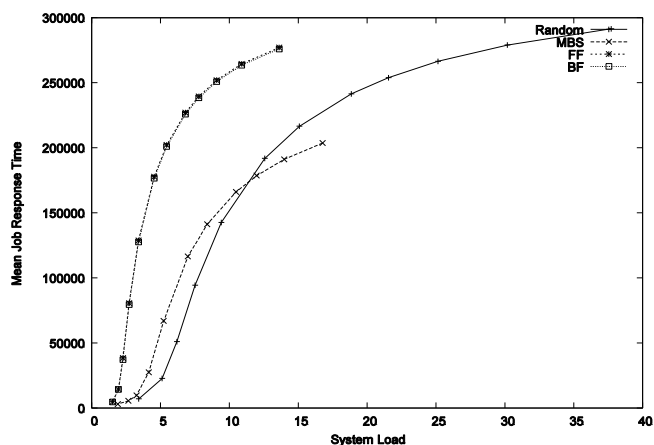


Fig. 1. Example Mean job response time (MJRT) vs. system load in multiple allocation strategies under the FCFS scheduling mechanism and all-to-all communication pattern (message count = 80)

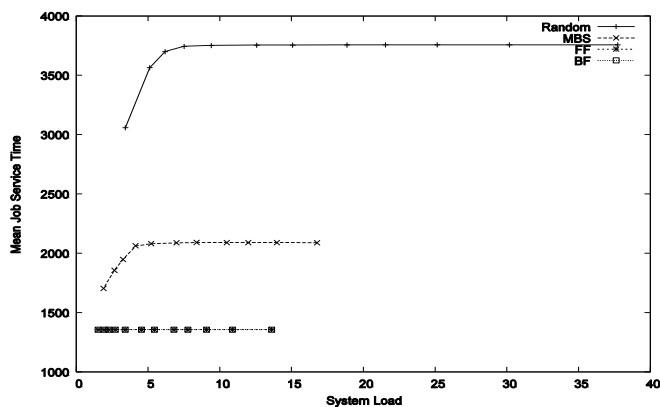


Fig. 2. Mean job service time vs. system load in multiple allocation strategies under the FCFS scheduling mechanism and all-to-all communication pattern (message count = 80)

This is because contiguous allocation strategies require that allocated processing units to be contiguous. The random allocation strategy showed less job response time compared to MBS at high system loads because that strategy randomly allocates the required processors.

Observation 2 (figure 2): Noncontiguous allocation strategies showed higher MJST than contiguous allocation strategies. This is because non-contiguous allocation strategies disperse the set of processors allocated to the parallel job when contiguous allocation fails. This increases the amount of time required for intra-process communication.

Observation 3 (figure 2): The MJST is higher when applying the random allocation strategy compared to the case when applying the MBS allocation strategy.

The random allocation strategy usually causes parallel jobs to be more dispersed compared to the MBS strategy. The MBS strategy is partially contiguous; it tries to maintain some level of contiguity between allocated processors.

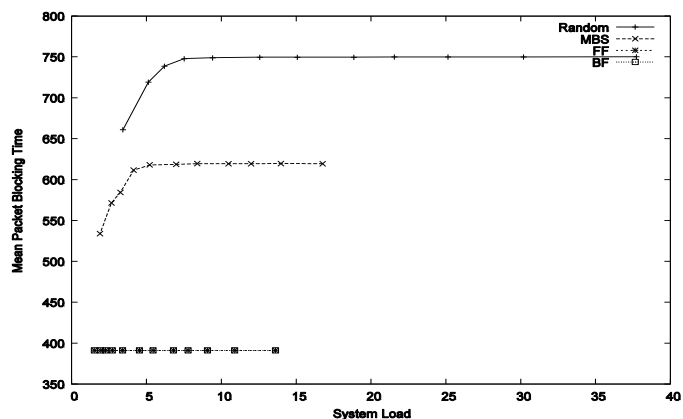


Fig. 3. Mean packet blocking time vs. system load in multiple allocation strategies under the FCFS scheduling mechanism and all-to-all communication pattern (message count = 80)

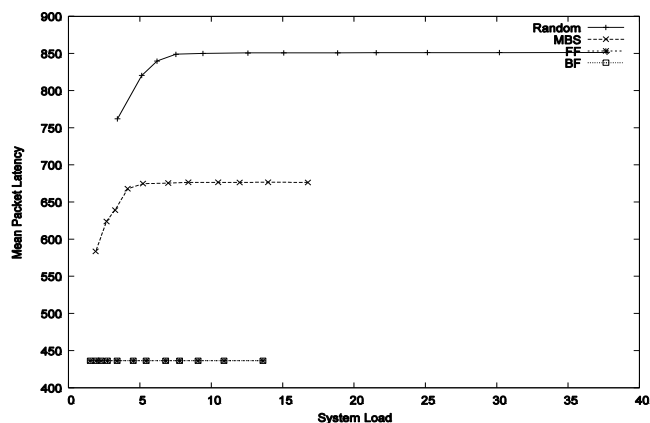


Fig. 4. Mean packet latency vs. system load in multiple allocation strategies under the FCFS scheduling mechanism and all-to-all communication pattern (message count = 80)

Non-contiguous allocation strategies causes more disperse to allocated parallel jobs producing longer distances between the sources and the destinations of messages. Since each message is blocked for sometime at each intermediate node

from source to destination, the mean packet blocking time increases as the number of these intermediate nodes increases. The mean packet latency also increases accordingly.

Observation 5 (figure 5): Contiguous allocation strategies showed low PSU values compared to noncontiguous allocation strategies. For example, at relatively high system loads, the random allocation strategy produces the highest percent system utilization of all experimented allocation strategies in figure 15. This is because it simply any idle processor from the list of free processors. The MBS strategy also allocates processors in a noncontiguous manner. Both strategies successfully allocate the parallel job in hand as long as enough number of free nodes is available. The FF and BF both forces the parallel job to wait if no enough contiguous set of processors is available in the system. This, in turn, may produce high waiting time for relatively large parallel requests.

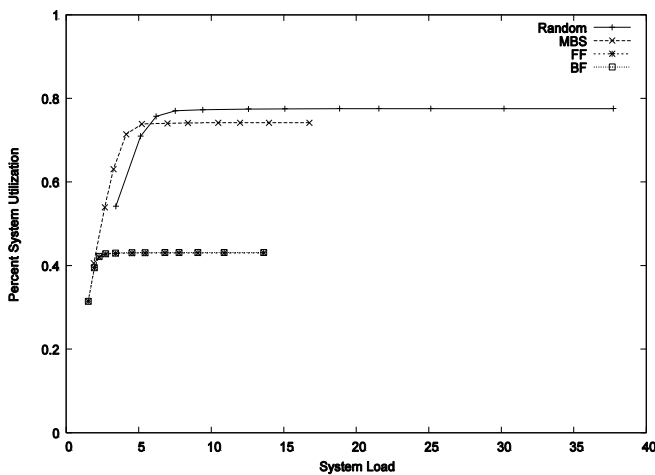


Fig. 5. Percent system utilization vs. system load in multiple allocation strategies under the FCFS scheduling mechanism and all-to-all communication pattern (message count = 80)

Figures 6 through 20 shows the relationships between the system load from one and, from the other side, (i) the mean job response time, (ii) the mean job service time, (iii) the mean packet blocking time, (iv) the mean packet latency, and (v) the percent system utilization. The impact of multiple communication patterns is studied next. Those patterns are: all-to-all, one-to-all, random, Fast Fourier Transform (FFT), and the NAS multigrid benchmark communication patterns on the key performance parameters of the multiprocessor system.

Figures 6 through 10 are specific for the contiguous allocation strategies. The authors use the BF allocation strategy as a representative for this group of allocation strategies.

Figures 11 through 15 are specific for the noncontiguous allocation strategies. The authors use the MBS allocation strategy as a representative for this group of allocation strategies.

Figures 16 through 20 are again specific for the noncontiguous allocation strategies. However, The authors use the BGP-BF allocation strategy as a representative for this

group of allocation strategies. In these experiments, the partitioning bound that is applied is equal to 4.

Observation 4 (figures 3 and 4): noncontiguous allocation strategies showed higher MPBT and MPL values than contiguous allocation strategies. This observation can be explained the same way as observations 3 and 4. Observation 9 (figures 6 and 7): One-to-all, All-to-all and random communication patterns produces the highest MJRT and MJST of all tested communication patters.

Next is the list of testbed communication patterns sorted starting with the one caused the highest MJRT: one-to-all, all-to-all, random, NAS-multigrid, FFT, and DQBT. This can be explained as follows, one-to-all and all-to-all communication patterns generate large number of messages to be transmitted over the interconnection network. This increases the service time of allocated parallel jobs and, thus, forces unallocated parallel jobs to wait more in the ready queue of the system. This, in turn, increases the mean job response time.

Figures 8 and 9 show the effect of communication pattern on the MPBL and MPL values of the system in the case of applying the BF allocation strategy. Notice the behavior of the system is quiet strange and that the figure is not directly conclusive. This is because that the number of jobs served during the simulator is relatively low because of the following factors:

(i)The allocation strategy applied in contiguous. This reduces the number of allocated jobs due to the condition of contiguity.

(ii) The scheduling mechanism applied is the FCFS. This may prevent many parallel jobs from being allocated if a relatively big parallel job is residing at the head of the ready queue of the system. Based on the above two factors, the authors believe that the behavior of the system is not clear in the figures 8, 9 and 10. To remedy this problem and to better capture the impact of communication pattern type. The authors next examines applying a non-contiguous allocation strategy.

Next, the authors present their observations on these figures.

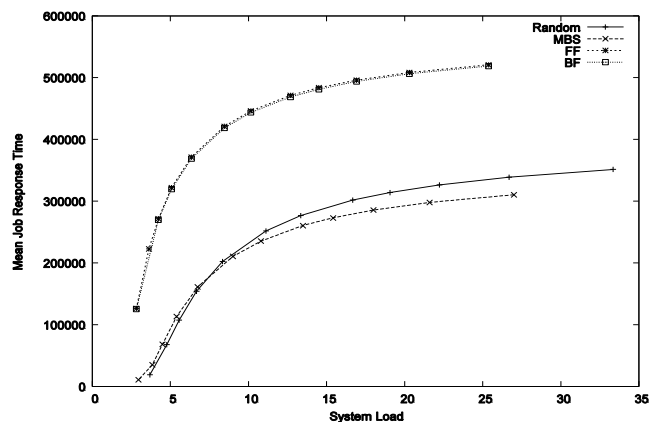


Fig. 6. Mean job response time vs. system load in Best Fit allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

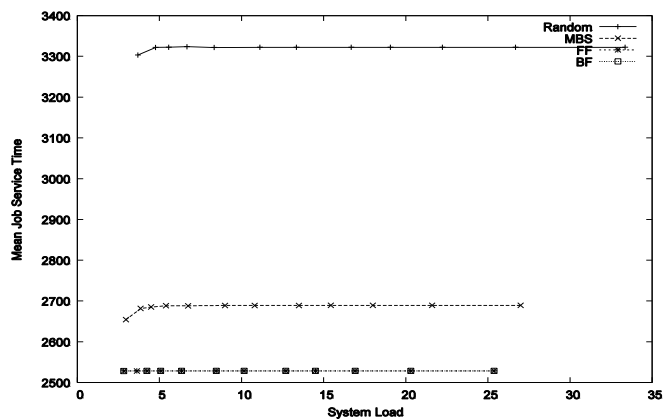


Fig. 7. Mean job service time vs. system load in Best Fit allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

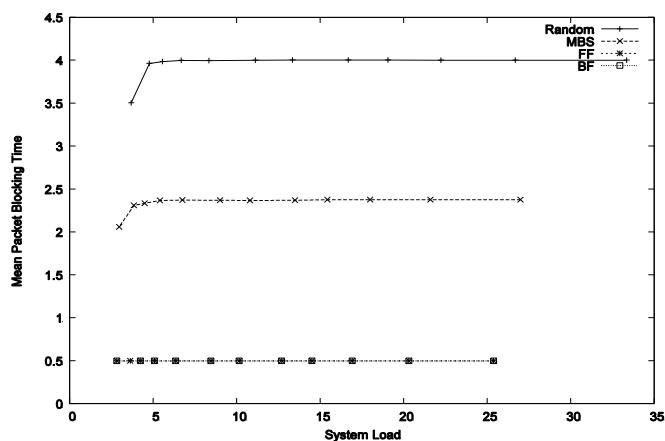


Fig. 8. Mean packet blocking time vs. system load in Best Fit allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

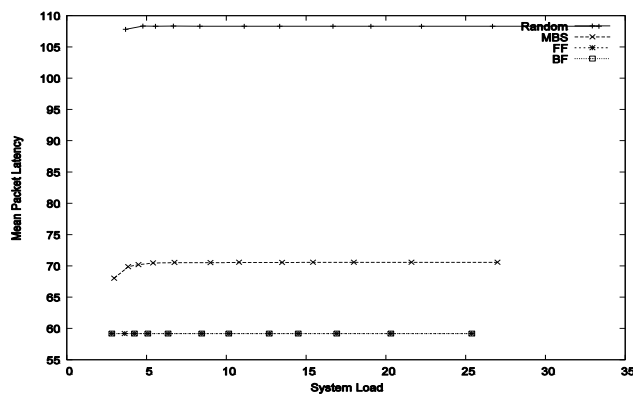


Fig. 9. Mean packet latency vs. system load in Best Fit allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 0)

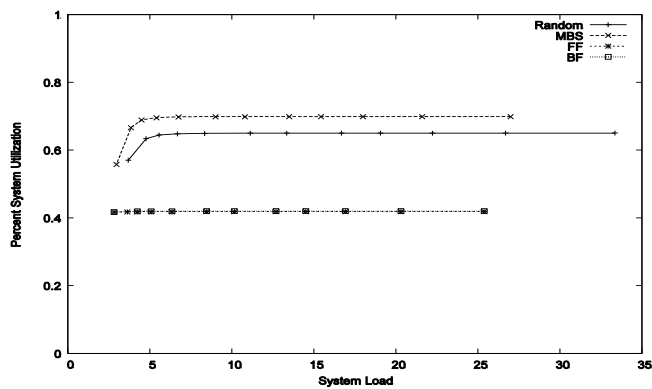


Fig. 10. Percent system utilization vs. system load in Best Fit allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

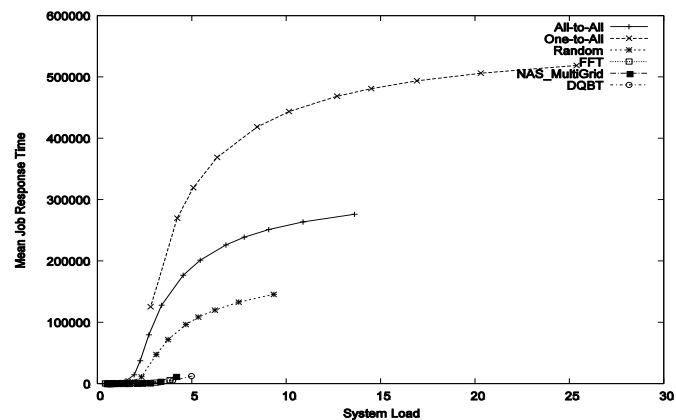


Fig. 11. Mean job response time vs. system load in MBS allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

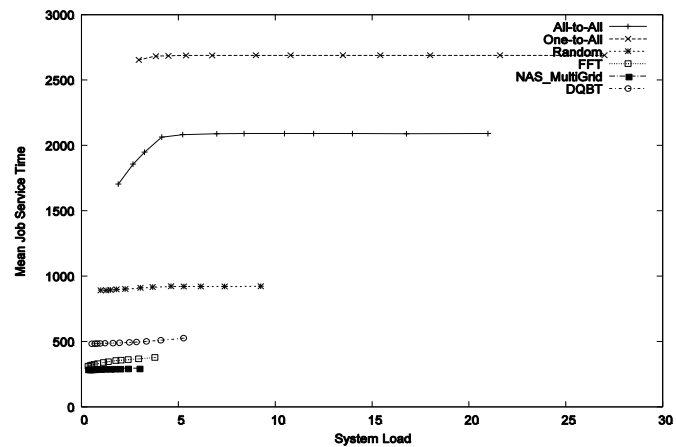


Fig. 12. Mean job service time vs. system load in MBS allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

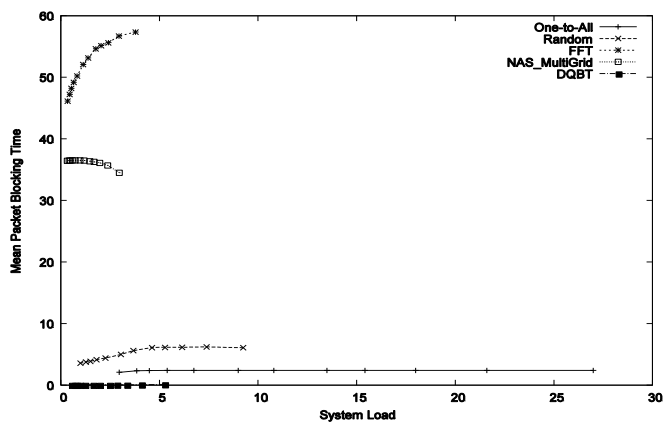


Fig. 13. Mean packet blocking time vs. system load in MBS allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

Figures 13 and 14 plot the MPBT, the MPL and the PSU vs system load when using the MBS allocation strategy and multiple communication patterns.

Observation 11 (figures 8, 9, 10, 13, 14, and 25): the behavior of the parallel system can be significantly affected by the communication pattern of the parallel jobs being allocated.

Figures 11 through 15 are specific for the noncontiguous allocation strategies. The authors use the MBS allocation strategy as a representative for this group of allocation strategies.

Figures 11 and 12 showed similar results to figures 6 and 7 that have been explained before (with the BF allocation strategy). One difference can be observed.

Figures 16 to 20 show the relationships between the system load from one side, and, from the other side, (i) the mean job response time, (ii) the mean job service time, (iii) the mean packet blocking time, (iv) the mean packet latency, and (v) the percent system utilization., that's focus to the BGP-BF allocation strategy.

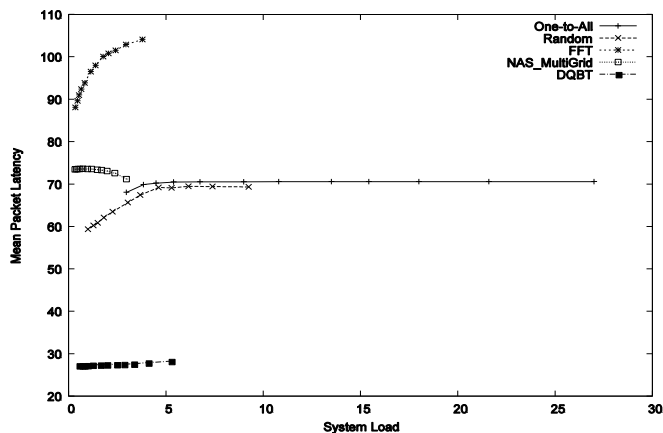


Fig. 14. Mean packet latency vs. system load in MBS allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

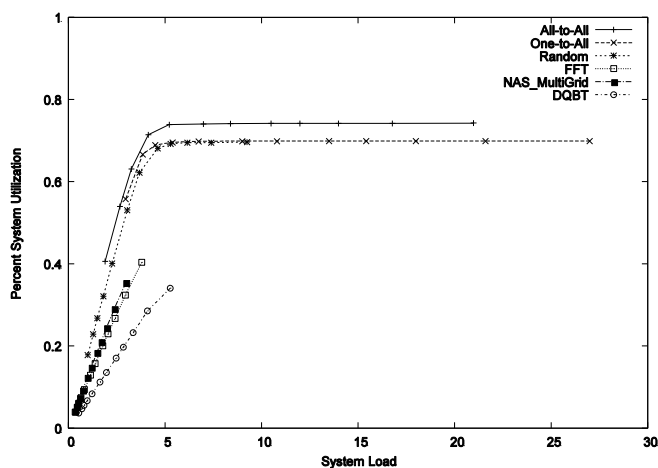


Fig. 15. Percent system utilization vs. system load in MBS allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

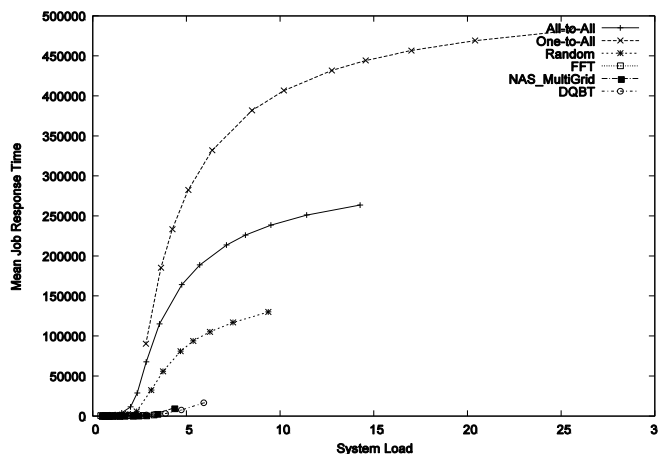


Fig. 16. Mean job response time vs. system load in the BGP-BF (partitioning bound = 4) allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

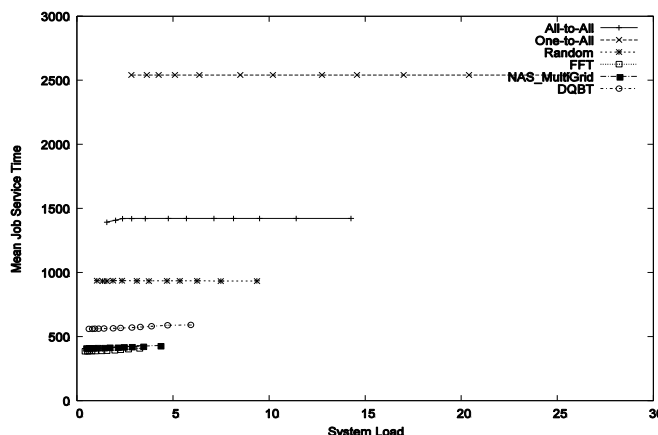


Fig. 17. Mean job service time vs. system load in the BGP-BF (partitioning bound = 4) allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)



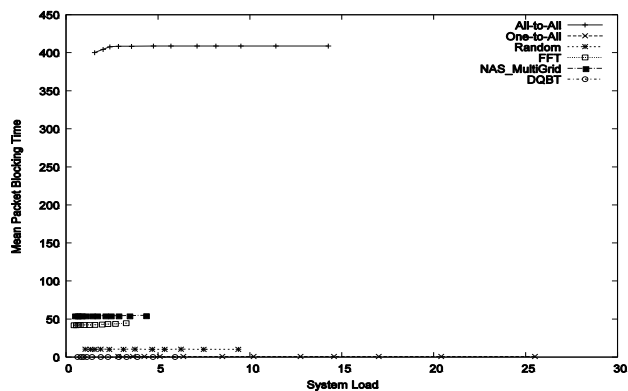


Fig. 18. Mean packet blocking time vs. system load in the BGP-BF (partitioning bound = 4) allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

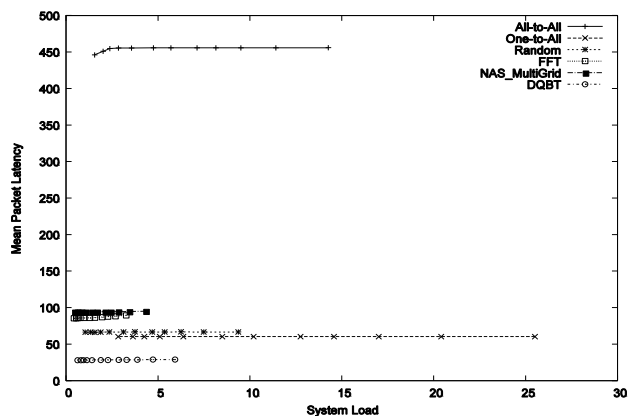


Fig. 19. Mean packet latency vs. system load in the BGP-BF (partitioning bound = 4) allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80). Conclusion

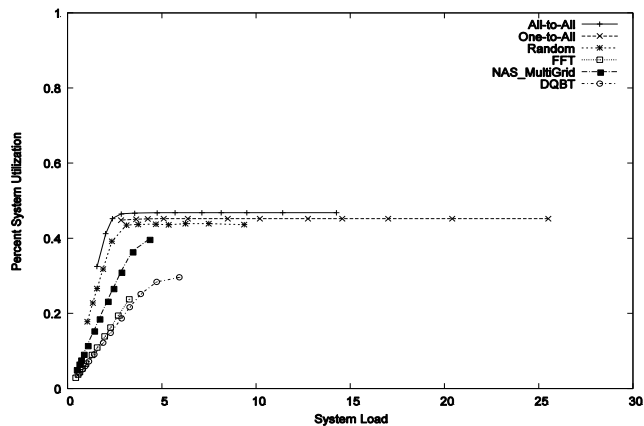


Fig. 20. Percent system utilization vs. system load in the BGP-BF (partitioning bound = 4) allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

Observation 12 (figures 10, 15 and 20): the behavior of the parallel system can be significantly affected by the communication pattern of the parallel jobs being allocated.

For instance, the random and the one-to-all communication patterns produced similar PSU value when applying the MBS allocation strategy. However, the one-to-all allocation strategy

outperformed the random communication pattern when using the BF and the BGP-BF strategy. Next, the authors study the impact of the partitioning bound of the BGP-BF allocation strategy on the different system performance parameters. This will be in figures 11 through 20. Figures 11 to 15 are specific for the all-to-all communication pattern. While figures 16 to 20 are specific for the communication pattern one-to-all.

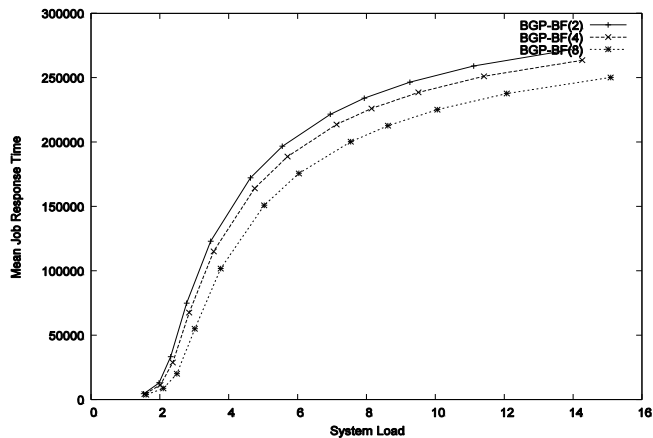


Fig. 21. Mean job response time vs. system load in the BGP-BF (partitioning bounds = 2, 4, and 8) allocation strategy under the FCFS scheduling mechanism and all-to-all communication pattern (message count = 80)

Observation 13 (Figure 21): In general, when the bounded is high the mean job response time is less.

For example, the BGP-BF (2) produced the highest MJRT value and followed by BGP-BF (4) and BGP-BF (8).

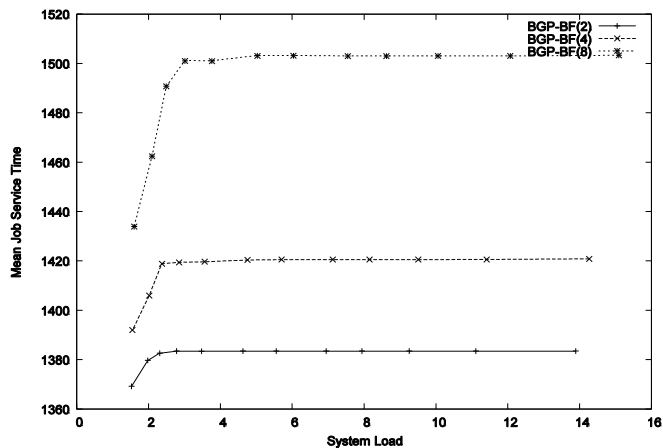


Fig. 22. Mean job response time vs. system load in the BGP-BF (partitioning bounds = 2, 4, and 8) allocation strategy under the FCFS scheduling mechanism and all-to-all communication pattern (message count = 80)

Figure 22 represents the mean job service time and system load in BGP-BF (partitioning bounds=2, 4 and 8) allocation strategy.

Observation 14 (figure 22): When increasing the partitioning bounded, the mean job service time increases also.

This is because when increasing the partitioning bound, the execution time increases because the level on non-contiguity increases. The results in an increase in the service time of the job.

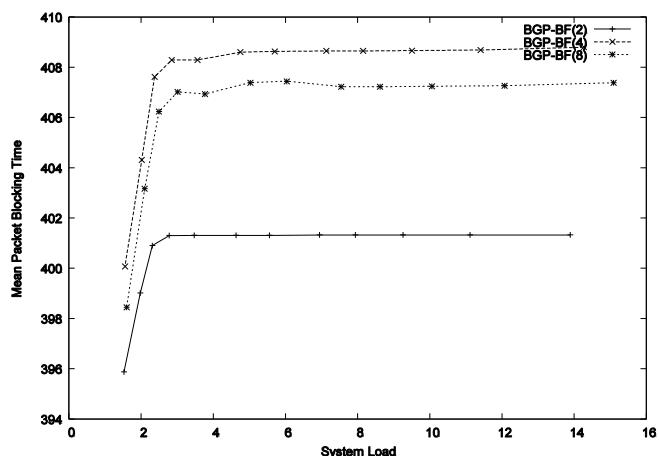


Fig. 23. Mean packet blocking time vs. system load in the BGP-BF (partitioning bounds = 2, 4, and 8) allocation strategy under the FCFS scheduling mechanism and all-to-all communication pattern (message count = 80)

Figures 23, 24 plot the relation between the mean packet blocking time and the mean packet latency (y-axis) and the system load (x-axis) when applying the BGP-BF (partitioning bounds = 2, 4, and 8) allocation strategy under the FCFS scheduling mechanism and all-to-all communication pattern.

Observation 15 (figures 23 and 24): In general, when the partitioning bound is high the MPBT and the MPL is high.

This is because when the bound is high the time that message packets spend blocked in network buffers, waiting for access to their next channel is high.

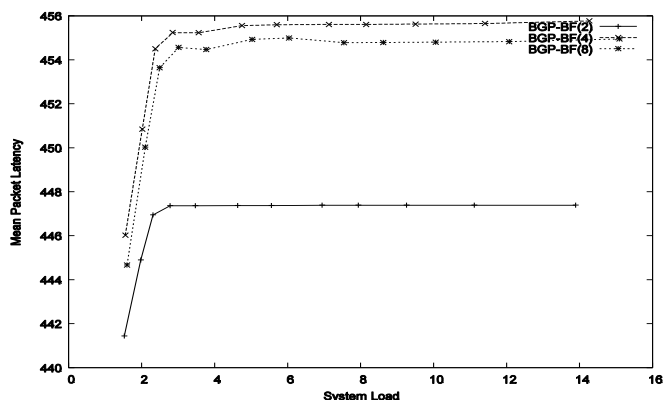


Fig. 24. Mean packet latency vs. system load in the BGP-BF (partitioning bounds = 2, 4, and 8) allocation strategy under the FCFS scheduling mechanism and all-to-all communication pattern (message count = 80)

The figure 24 that represent the relation between mean packet latency and system load in the BGP-BF "partitioning bounds=2, 4 and 8" allocation strategy under the FCFS scheduling mechanism and all-to-all communication pattern.

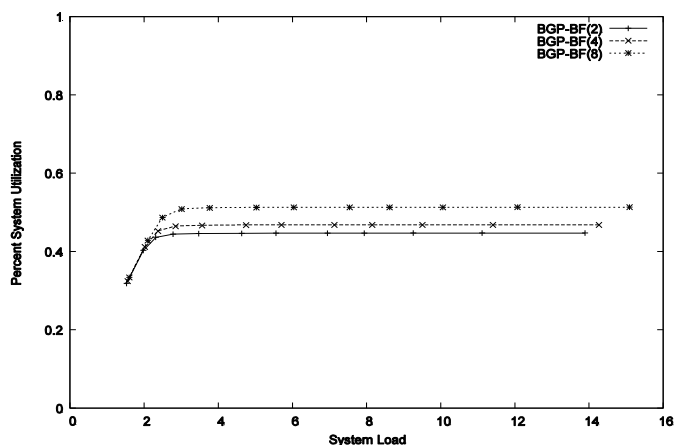


Fig. 25. Percent system utilization vs. system load in the BGP-BF (partitioning bounds = 2, 4, and 8) allocation strategy under the FCFS scheduling mechanism and all-to-all communication pattern (message count = 80)

Figure 25 plots the percent system utilization and system load in the BGP-BF (partitioning bounds = 2, 4, and 8) allocation strategy under the FCFS scheduling mechanism and all-to-all communication pattern.

Observation 16 (figure 25): when the partitioning bound increase so the percent system utilization increases.

BGP-BF (8) showed the highest PSU values, followed by the BGP-BF (4), BGP-BF (2) respectively. This because it's when the bound is higher the processors of the multicomputer system is utilized more efficiently.

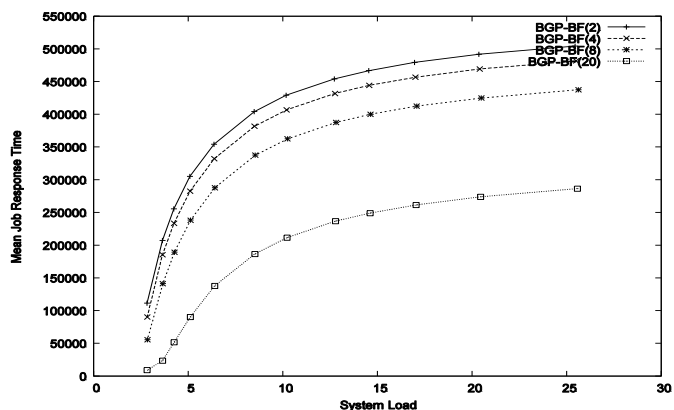


Fig. 26. Mean job response time vs. system load in the BGP-BF (partitioning bounds = 2, 4, 8, and 20) allocation strategy under the FCFS scheduling mechanism and one-to-all communication pattern (message count = 80)

Figures 26 through 30 (one-to-all communication pattern) gives similar observations to the figures 36 through 40 (all-to-all communication pattern).

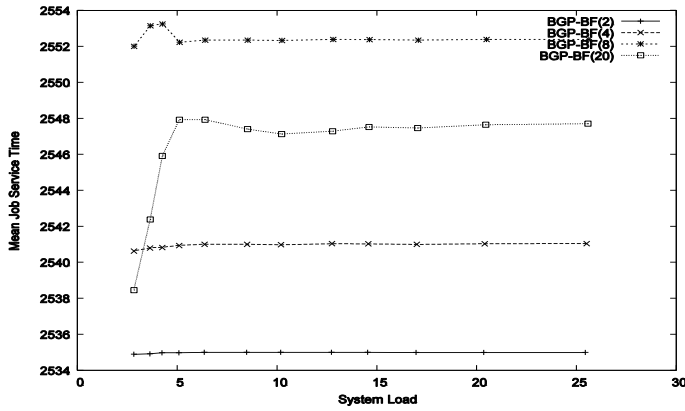


Fig. 27. Mean job service time vs. system load in the BGP-BF (partitioning bounds = 2, 4, 8, and 20) allocation strategy under the FCFS scheduling mechanism and one-to-all communication pattern (message count = 80)

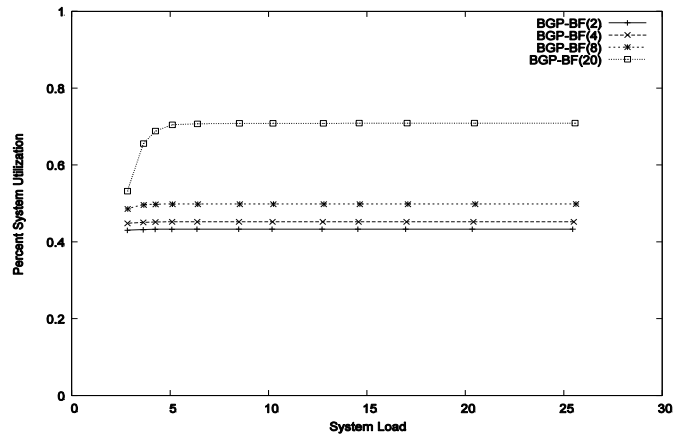


Fig. 30. Percent system utilization vs. system load in the BGP-BF (partitioning bounds = 2, 4, 8, and 20) allocation strategy under the FCFS scheduling mechanism and one-to-all communication pattern (message count = 80)

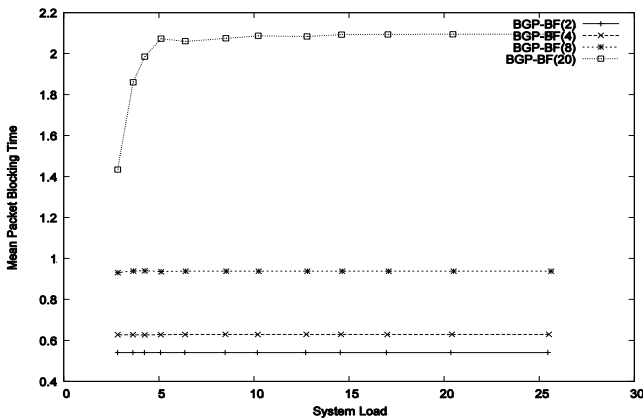


Fig. 28. Mean packet blocking time vs. system load in the BGP-BF (partitioning bounds = 2, 4, 8, and 20) allocation strategy under the FCFS scheduling mechanism and one-to-all communication pattern (message count = 80)

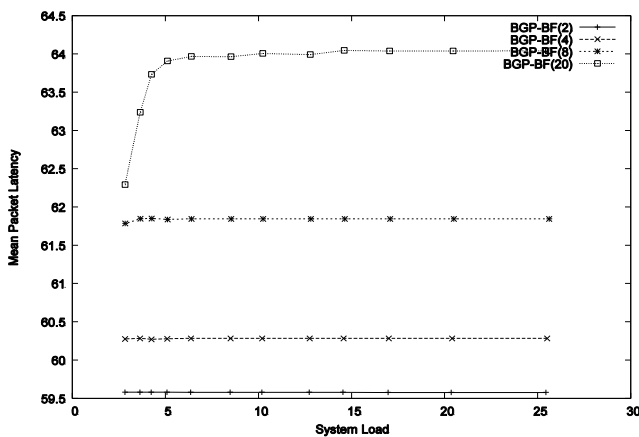


Fig. 29. Mean packet latency vs. system load in the BGP-BF (partitioning bounds = 2, 4, 8, and 20) allocation strategy under the FCFS scheduling mechanism and one-to-all communication pattern message count = 80)

Figures 31 through 35 represent the relation between the system load and the following system performance parameters: the mean job response, the mean job service time, the mean packet blocking time, mean packet latency and the percent system utilization. Here, the authors focus on the BGP-BF (partitioning bound=8) allocation strategy under the FCFS scheduling mechanism and multiple communication patterns such as all-to-all, one-to-all, random, NAS\_multigrid, DQBT.

Notice that how the key system parameters; the MJRT, MJST, MPBT, MPL and the PSU change over changing the communication pattern of the parallel jobs being scheduled.

The above discussion and set of observations clearly shows that, depending on the performance parameter of most interest, the system can dynamically modify the partitioning bound of the BGP allocation strategy.

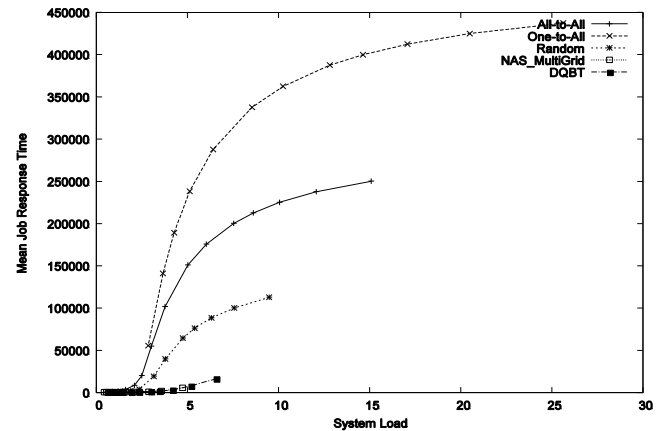


Fig. 31. Mean job response time vs. system load in the BGP-BF (partitioning bound = 8) allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

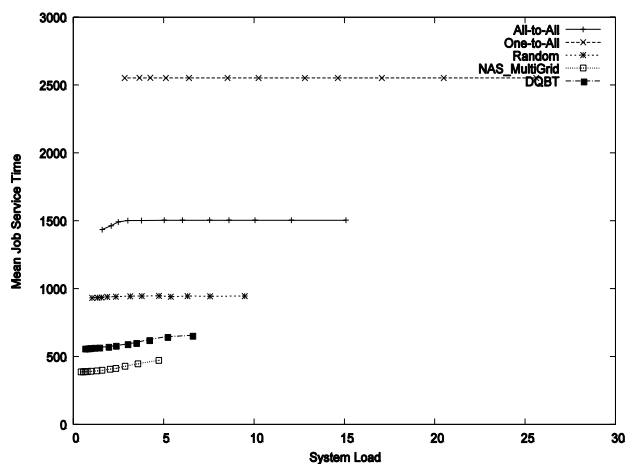


Fig. 32. Mean job service time vs. system load in the BGP-BF (partitioning bound = 8) allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

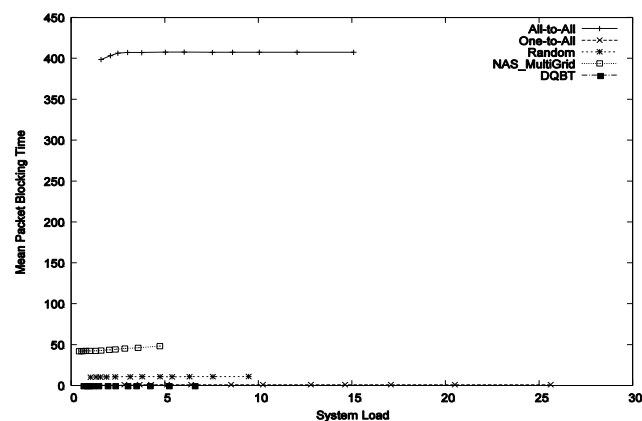


Fig. 33. Mean packet blocking time vs. system load in the BGP-BF (partitioning bound = 8) allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

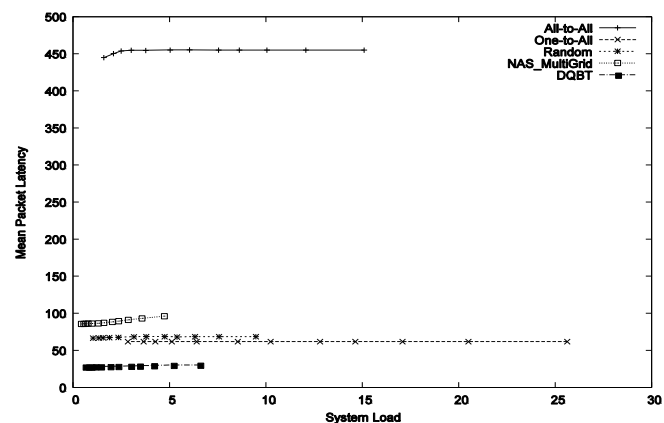


Fig. 34. Mean packet latency vs. system load in the BGP-BF (partitioning bound = 8) allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

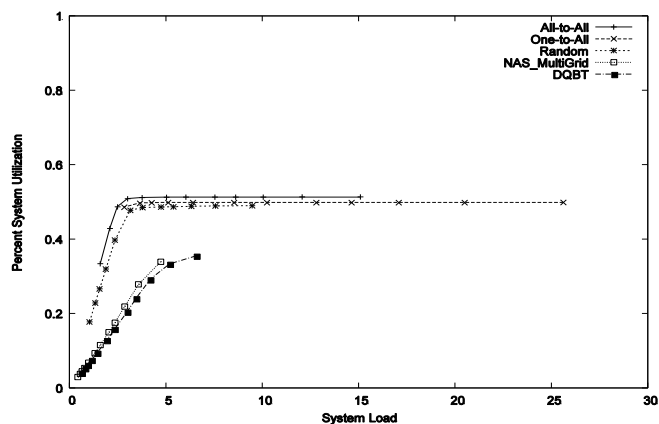


Fig. 35. Percent system utilization vs. system load in the BGP-BF (partitioning bound = 8) allocation strategy under the FCFS scheduling mechanism and multiple communication patterns (message count = 80)

## V. CONCLUSION

A key performance factor that can highlight the difference between allocation strategies is the amount of communication conducted between the parallel jobs to be allocated. In this paper, the authors show that's the type and pattern of communication can affect the performance of these strategies, compared to the communication pattern that are usually used in literature to evaluate processor allocation strategies, the authors examined wider range of communication patterns in the current work. Other works consider only two types of communication patterns; those are the one-to-all and all-to-all patterns. The authors found that the communication behavior of the parallel jobs being allocated can have a significant impact on the performance of the processor allocation strategy being applied. This observation is correct for both, contiguous and non contiguous strategies.

## REFERENCES

- [1] Ababneh, I., "An efficient free-list sub-mesh allocation scheme for two-dimensional mesh-connected multicomputers", *Journal of Systems and Software*, vol. 79, no. 8, Elsevier Science Inc., New York, NY, USA, August 2006, pp. 1168-1179.
- [2] Alqadi, R. and M. Khammash. "An Efficient Parallel Gauss-Seidel Algorithm for the Solution of Load Flow Problems". *The International Arab Journal of Information Technology*. Vol. 4, No. 2. April 2007.
- [3] Bani-Ahmad, S., "Sub-mesh Allocation in 2D-Mesh Multicomputer: Partitioning at the Longest Dimension of Requests". *Proceedings of the Fourth International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP 2010)*. October 25-30, 2010, Florence, Italy.
- [4] Bani-Ahmad, S., "Bounded Grand-Request-Based Allocation Strategies in 2D Mesh-based Multicomputers". *International Journal of Digital Content Technology and its Applications*. Volume 5, Number 1, January 2011.
- [5] Bani-Ahmad, S., "Comparative Evaluation of Request-Partitioning-Based Processor Allocation Strategies in 2D Mesh-based Multicomputers". *International Journal of Computer Applications (0975-8887)*. Volume 26, Number 7, July 2011.
- [6] Bani-Ahmad, S., "Processor Allocation with Reduced Internal and External Fragmentation in 2D Mesh-based Multicomputers". *Journal of Applied Sciences*. Volume 11, Issue: 6, 2011, pp 943-952.

- [7] Bani-Mohammad, S.; Ould-Khaoua, M.; Ababneh, I., and Machenzie, L., "Non-contiguous Processor Allocation Strategy for 2D Mesh Connected Multicomputers Based on Sub-meshes Available for Allocation", Proceedings of the 12th International Conference on Parallel and Distributed Systems (ICPADS'06), vol. 2, IEEE Computer Society Press, USA, 2006, pp. 41-48.
- [8] Chuang, P. J. and N.F.Tzeng, "Allocating Precise sub-mesh in Mesh-Connected Systems" IEEE Trans. on Parallel and Distributed Systems, pp.211-217, Feb, 1992.
- [9] Ding, J. and L. N. Bhuyan, "An Adaptive Sub-mesh Allocation Strategy for Two-Dimensional Mesh Connected Systems", Proceedings of International Conference on Parallel Processing, Vol.11, pp.193-200, Aug.1993.
- [10] Kumar, V., Grama A., Gupta A., and Karypis G.. "Introduction To Parallel Computing", The Benjamin/Cummings publishing Company, Inc., Redwood City, California, 2003.
- [11] Lo, V., Windisch K., Liu W., and Nitzberg B.. "Non-contiguous processor allocation algorithms for mesh-connected multicomputers", IEEE Transactions on Parallel and Distributed Systems, vol. 8, no. 7, IEEE Press, Piscataway, NJ, USA, July 1997, pp. 712-726.
- [12] ProcSimity V4.3 User's Manual, University of Oregon, [ftp://ftp.cs.uoregon.edu/pub/lo/procsim\\_man.ps.gz](ftp://ftp.cs.uoregon.edu/pub/lo/procsim_man.ps.gz), 1997.
- [13] Seydim A. Y., "Wormhole Routing in Parallel Computers", school of Engineering and Applied Sciences, Southern Methodist University, May 1998.
- [14] Srinivasan, T., Seshadri J., Chandrasekhar A., and Jonathan J.. "A Minimal Fragmentation Algorithm for Task Allocation in Mesh-Connected Multicomputers", Proceedings of IEEE International Conference on Advances in Intelligent Systems – Theory and Applications – AISTA 2004 in conjunction with IEEE Computer Society, ISBN 2-9599-7768-8, IEEE Press, Luxembourg, Western Europe, 15-18 Nov 2004.
- [15] Suzaki, K., Tanuma H., Hirano S., Ichisugi Y., Connelly C., and Tsukamoto M.. "Multi-tasking Method on Parallel Computers which Combines a Contiguous and Non-contiguous Processor Partitioning Algorithm", Proceedings of the Third International Workshop on Applied Parallel Computing, Industrial Computation and Optimization, Springer-Verlag, UK, 1996, pp. 641-650.
- [16] Windisch, K.; Miller, J. V.; and Lo, V., "ProcSimity: an experimental tool for processor allocation and scheduling in highly parallel systems", Proceedings of the Fifth Symposium on the Frontiers of Massively Parallel Computation (Frontiers'95), IEEE Computer Society Press, Washington, USA, 6-9 Feb 1995, pp. 414-421.
- [17] Yoo, B. S. and Das, C. R.. "A Fast and Efficient Processor Allocation Scheme for Mesh-Connected Multicomputers", IEEE Transactions on Parallel & Distributed Systems, vol. 51, no. 1, IEEE Computer Society, Washington, USA, January 2002, pp. 46-60.
- [18] Zhu, Y. H., "Efficient Processor Allocation Strategies for Mesh-Connected Parallel Computers", Journal of Parallel and Distributed Computing, 16, pp 211-337, 1992
- [19] Massey, G. & McClintock, J.. "Packet Size Matters in IP Transport", Copyright 2006 IMAS Publishing (USA), Inc.
- [20] David Bailey, Tim Harris, William Saphir, Rob Van der Wijngaart, Alex Woo, Maurice Yarrow (1995), "The NAS Parallel Benchmarks 2.0", Report NAS-95-020, December, 1995.
- [21] F. Shimojo, R. K. Kalia, A. Nakano, and P. Vashishta, "embedded divide-and-conquer algorithm on hierarchical real-space grids: parallel molecular dynamics simulation based on linear-scaling density functional theory," Comput. Phys. Commun. 167, 151.
- [22] Rolf Riesen, "Communication patterns", IEEE In Workshop on Communication Architecture for Clusters CAC'06, Rhodes Island, Greece, April 2006.
- [23] Singh, K. & Kilarikar, N.. "Tcp Flow Control Mechanisms For Single Packet Loss", paper presented in the Proceedings of the National Conference "NCNTE-2012" at Fr. C.R.I.T., Vashi, Navi Mumbai, Feb. 24-25, 2012. 1
- [24] K. Wendy Tang, Member, IEEE, and Sanjay A. Padubidri, " Diagonal and Toroidal Mesh Networks", IEEE TRANSACTIONS ON COMPUTERS ,VOL.43 ,NO.7, July, 1994
- [25] Li, K. and K.H Cheng, "A Two-Dimensional Buddy System for Dynamic Resource Allocation in a Partitionable Mesh Connected System", Journal of Parallel and Distributed Computing, 12, pp.79-83, 1991.

# ECG Signal Compression Using the High Frequency Components of Wavelet Transform

Surekha K.S

(Research Student, Sinhgad College of Engg., Pune,India)  
Assoc. Professor, Department of E&TC  
Army Institute of Technology, Dighi, Pune, India

B. P. Patil

Principal  
Army Institute of Technology, Dighi,  
Pune, India

**Abstract**—Electrocardiography (ECG) is the method of recording electrical activity of the heart by using electrodes. In ambulatory and continuous monitoring of ECG, the data that need to be handled is huge. Hence we require an efficient compression technique. The data also must retain the clinically important features after compression. For most of the signals, the low frequency component is considered as most important part of the signal. In wavelet analysis, the approximation coefficients are the low frequency components of the signal. The detail coefficients are the high frequency components of the signal. Most of the time the detail coefficients (high frequency components) are not considered. In this paper, we propose to use detail coefficients of Wavelet transform for ECG signal compression. The Compression Ratio (CR) of both the approximation and detail coefficients are compared. Threshold based technique is adopted. The Threshold value helps to remove the coefficients below the set threshold value of coefficients. Experiment is carried out using different types of Wavelet transforms. MIT BIH ECG data base is used for experimentation. MATLAB tool is used for simulation purpose. The novelty of the method is that the CR achieved by detail coefficients is better. CR of about 88% is achieved using Sym3 Wavelet. The performance measure of the reconstructed signal is carried out by PRD.

**Keywords**—ECG; PRD; transform

## I. INTRODUCTION

ECG is used as an important signal which gives information about health of the heart. It also helps physicians to diagnose cardiac arrhythmias and different heart diseases. The heart rate can also be determined by calculating time between successive QRS complexes.

The different methods of electrocardiography [1, 2] are standard clinical ECG, Vector cardiogram and ECG signals are recorded continuously to monitor the abnormality.

Electrocardiogram signal is an important source of information for doctors in diagnosis purpose. It helps the doctors to diagnose heart abnormalities. Most of the time, due to continuous monitoring of ECG signals in 24 hour monitoring system and in ambulatory system, the data storage requirement increases. This in turn increases the storage cost. Instead of transmitting the stored data directly, ECG signal is compressed before transmission through common communication channels like phone line or mobile channel. The compression of ECG signal reduces the storage and transmission cost. The important factor to be considered in

compression of ECG signal is to obtain maximum data reduction. At the same time, the clinically important features of ECG signal must be preserved after reconstruction.

The ECG signal is shown in fig. 1. It consists of P, Q, R, S, T and U waves. The first positive wave is the P wave. The QRS complex consists of Q, R and S waves. The wave is produced by ventricular activation. The T wave is produced by ventricular repolarization and it is a smooth dome shaped. The U wave follows the T wave and precedes the P wave of the next cycle.

This paper is structured as follows: section 2 describes the lossy and lossless compression techniques; Section 3 describes the performance measure to validate the technique used in this research; Section 4 covers the methodology. Results and discussion are covered in section 5 and 6 respectively. Conclusion and references are the concluding sections.

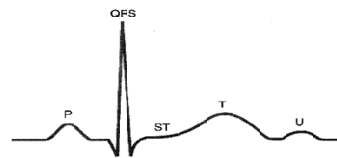


Fig. 1. The normal ECG

## II. LOSSY AND LOSSLESS COMPRESSION

The main focus of the research is on compression. The basic types of compression are Lossy and Lossless. The lossy compression accepts slight loss of data [3]. Lossless data compression allows the original data to be perfectly reconstructed from the compressed data. Lossy compression allows reconstruction only of an approximation of the original data. The lossy compression also improves compression ratio which in turn decreases the storage space and also the transmission cost. In lossy compression method, better compression is achieved by losing some information. When the decompression takes place, the resulting output data does not match with the original input data. If the loss of data is small, it is difficult to tell the difference [4]. When we consider text files, especially files containing computer programs, even if a single bit gets modified, it becomes worthless. In such cases, we require only lossless compression techniques. Higher compression ratio is achieved in Lossy compression method. Most of the compression techniques used for ECG is lossy.



In an ambulatory and continuous monitoring of data, huge data need to be handled. This leads to huge amount of data for storage and transmission. Hence the ECG data needs a proper compression technique, for data reduction [5-13, 20]. Various compression methods used are:

1) *Direct Method*: In the Direct method, time domain signal samples are analyzed. The examples of direct methods are TP (Turning Point), AZTEC (Amplitude Zone Time Epoch Coding), ASEC (Analysis by Synthesis ECG Compressor), CORTES (Coordinate Reduction Time Encoding System), FAN etc. [5].

2) *Transformational Method*: The technique uses transformation of the signal from one domain to another domain. The examples of transformation method are Discrete Cosine Transform, Fourier Transform, Wavelet Transform, Walsh Hadamard Transform, Fast Fourier Transform[17], Fractional Fourier transform[18, 19] etc.

3) *Parameter Extraction Methods*: The method uses a set of parameters extracted from the original signal and are used in the reconstruction process.

The advantage of direct method is its simplicity. In the transform, small transform coefficients become insignificant and can be removed. This way, a better compression ratio is achieved in transform technique. The Wavelet Transform technique is considered in this research and a brief description of the same is given below.

A. *The Wavelet Transform*

The wavelet transform is suitable for ECG signal compression because of its non-stationary nature. The wavelet transform is also capable of capturing both frequency and location in time information. The different types of Wavelet transforms are continuous wavelet transform (CWT), a wavelet series expansion, and a DWT [6, 14].

A DWT is the one for which the wavelets are discretely sampled. An important advantage of Wavelet Transform over Fourier transform is that it captures both frequency and location in time information.

The signal x is passed through a series of filters to obtain DWT of the signal. As a first step, the samples are passed through a low pass filter with impulse response g. This procedure leads to convolution of the input signal with the impulse response.

$$y[n] = (x * g)[n] = \sum_{k=-\infty}^{\infty} x[k]g[n - k] \tag{1}$$

Signal is then simultaneously passed through high pass filter with impulse response h. The outputs from high pass filter are called the detail coefficients. The outputs from the low pass filter are called the approximation coefficients. The filter outputs are then subsampled by 2. The block diagram of filter analysis is shown in fig. 2.

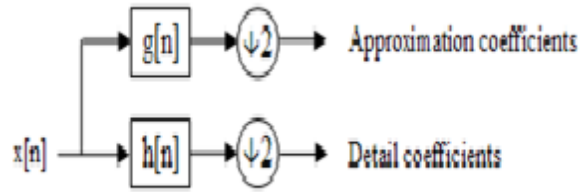


Fig. 2. Block diagram of filter analysis

The wavelet transform achieves higher frequency resolution by cascading. The cascading structure is shown in fig. 3. In this structure, the decomposition procedure is repeated. In this methodology, approximation coefficients are decomposed with high pass and low pass filters. After the filtering process, the down sampling is also repeated. Two level filter decomposition is shown in fig. 3.

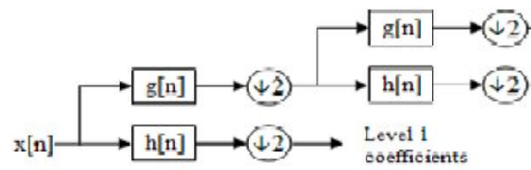


Fig. 3. A two level filter bank

From level 2, the further expansion takes place to higher levels. At each level the signal is decomposed into low and high frequencies.

The discrete wavelet transform is defined as:

$$f(t) = \sum C_{m,n} \psi_{m,n}(t) \tag{2}$$

III. PERFORMANCE MEASURE

The standard performance measure technique is used in this research to ascertain the usefulness of the compression method adopted. The standard methods are Percent Root mean square Difference (PRD), Compression Ratio [CR][16] and Peak Signal to Noise Ratio(PSNR).

The PRD is defined as the distortion between the original and the reconstructed signal.

$$PRD = \sqrt{\frac{\sum_{n=0}^{N-1} [x_0(n) - x_r(n)]^2}{\sum_{n=0}^{N-1} [x_0^2(n)]}} \times 100 \% \tag{3}$$

Here,  $x_0$  indicates the original data,  $x_r$  indicates the reconstructed data, and N represents the number of samples. The CR is defined as the ratio of number of bits in the original signal to the number of bits in the compressed signal.

$$CR = \frac{\text{Number of bits in the original signal}}{\text{Number of bits in the compressed signal}} \tag{4}$$

IV. METHODOLOGY

Different types of noise and interferences corrupt the ECG signal during recording [15]. As the ECG signal is susceptible to noise and interference, the signal is passed through various stages to remove the noise and interference. In the preprocessing stage, a High Pass filter is used to remove the base line wander noise and Band Pass filter is used to remove the power line interference. A threshold value is chosen in this method. The Threshold value helps to remove the coefficients below the set threshold value of coefficients. Decomposition of wavelet transform is restricted to the first level. The detail coefficients are used and the less important coefficients will be made zero. Compression ratio of the signal is increased with this threshold based method.

V. RESULTS AND IMPLEMENTATION

The ECG signal used for the testing purpose is taken from MITBIH data base [21]. The MATLAB tool is used for simulation. We have used a sampling frequency of 330 Hz. The results are tabulated for different Wavelets. Table 1 shows results for the Haar Wavelet. Table 2 shows the results for the Db2. Table 3 shows the result Sym3 Wavelet.

The different levels of threshold levels are used for testing purpose. The threshold is taken with reference to the R peak. The level 1 (0.05% of R peak is considered as the lowest level and level 3 (1% of R peak) is considered as the highest level. Level selection is as follows:

- 3- Highest level
- Level 1=0.05% of R peak.
- Level 2=0.5% of R peak
- Level 3=1% of R peak

TABLE I. CR & PRD COMPUTATION FOR HAAR WAVELET

Threshold level	CR Approx. coefficients	CR Detail coefficients	PRD Approx. coefficients	PRD Detail coefficients
Level 1	0.45	0.90	0.75	1.07
Level 2	1.01	13.12	0.75	1.07
Level 3	48.86	58.82	1.45	1.22

TABLE II. CR & PRD COMPUTATION FOR DB2 WAVELET

Threshold level	CR Approx. coefficients	CR Detail coefficients	PRD Approximation coefficients	PRD Detail coefficients
Level 1	0	2.7	1.2	1.3
Level 2	2.2	23.07	1.2	1.3
Level 3	50.22	77.82	1.9	0.8

TABLE III. CR & PRD COMPUTATION FOR SYM3 WAVELET

Threshold level	CR Approx. coefficients	CR Detail coefficients	PRD Approximation coefficients	PRD Detail coefficients
Level 1	0.42	4.52	1.11	0.93
Level 2	1.1	42.53	1.1	0.88
Level 3	50.22	88.23	1.76	0.36

The graphs for CR and PRD for different threshold levels are shown below. The graph of CR versus threshold level for Haar, db2 & Sym3 Wavelets are shown in fig 4, 5 & 6 respectively.

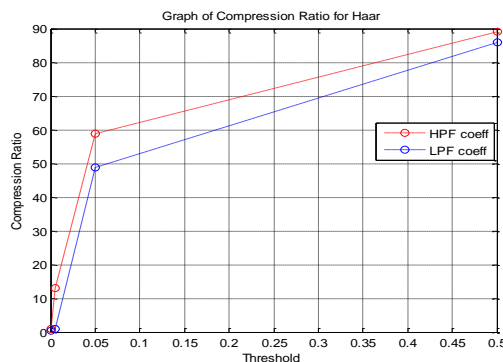


Fig. 4. The graph of CR versus threshold level for Haar wavelet

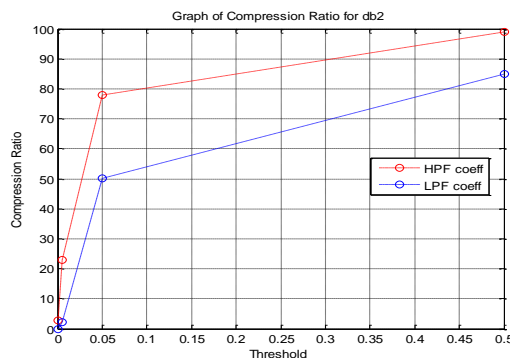


Fig. 5. The graph of CR versus threshold level for db2 Wavelet

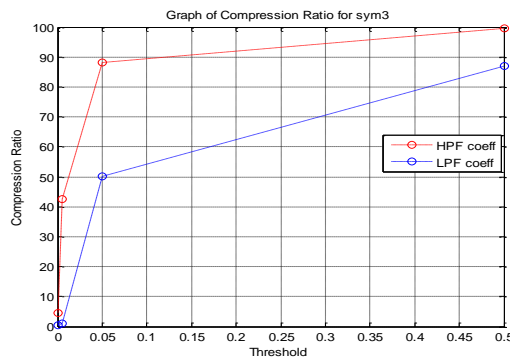


Fig. 6. The graph of CR versus threshold level for Sym3 wavelet

The graph of PRD versus threshold level for Haar, db2 & Sym3 are shown in fig 7, 8 and 9 respectively.

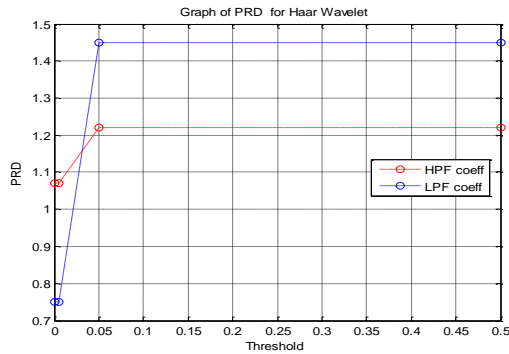


Fig. 7. The graph of PRD versus threshold level for Haar Wavelet

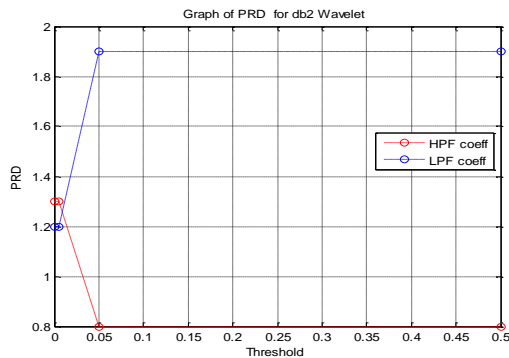


Fig. 8. The graph of PRD versus threshold level for db2 Wavelet

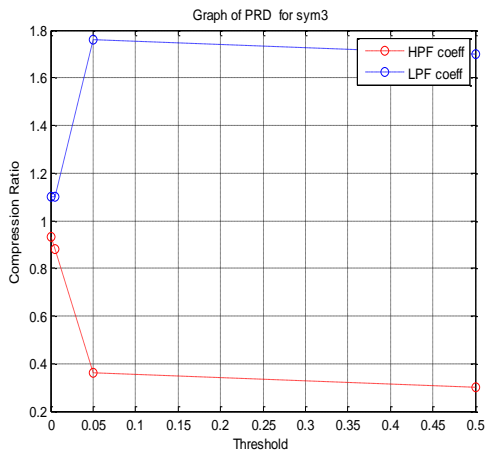


Fig. 9. The graph of PRD versus threshold level for Sym3 Wavelet

## VI. DISCUSSION

The main focus of our research to use the detail coefficients of wavelet transform for ECG signal compression. Various wavelet transforms like Haar, db2 and Sym3 are considered for experimentation. The performance evaluation is carried out to test the effectiveness of the compression technique. As shown in Table 3, the Sym3 wavelet shows better CR among all the three wavelets used. The PRD value indicates reconstruction quality of the signal after compression. The PRD value obtained in the experimentation

are within the limits. The threshold levels have significant impact on the CR value. Results obtained for the threshold level 1 shows minimum value of CR. As the threshold value increases, the CR also increases.

## VII. CONCLUSION AND FUTURE SCOPE

The methodology used in our research to investigate the usefulness of detail coefficients has given better results. The results also show that the compression of ECG signal is possible using detail coefficients. Detail coefficients show significant improvement in the CR value compared to the approximation coefficients. The Sym3 wavelet has done well among the different wavelets compared here. Proper selection of threshold level plays an important role in the reconstructed signal. As the level of threshold increases, the PRD also increases. Due to this, distortion may occur in the reconstructed signal.

The results with the change in threshold level are also highlighted in the result table. It is possible to achieve better compression ratio & PRD by adjusting threshold value.

Further improvement in compression ratio may be possible by using efficient encoding techniques.

## ACKNOWLEDGMENT

The authors would like to thank the Principal, Head of Department and PhD. coordinator, Sinhgad college of Engineering, for their support & guidance to publish this paper.

The authors would like to thank Director and Joint Director, AIT for the support and guidance to carry out research in Institute.

## REFERENCES

- [1] Wills J Tomkins, "Biomedical Digital Signal Processing", PHI 10<sup>th</sup> edition.
- [2] R S Khandpur, "Handbook Of Biomedical Instrumentation", TMH, 6<sup>th</sup> edition
- [3] Mark Nelson, Jean Loup Gailly. "The Data Compression Book",BPB,2nd Edition,2009
- [4] David Solomon, "Data Compression The Complete Reference",Springer International Edition,4th Edition,2011
- [5] S. S. Jalaeddine, C. Hutchens, R. Strattan and W. Coberly, "ECG Data Compression Techniques- A Unified Approach", IEEE Trans. on Biomedical Engineering, Vol. 37, pp. 329-343, 1990.
- [6] Mohammad Pooyan, Ali Taheri, Morteza Moazami-Goudarzi, ImanSaboori, "Wavelet Compression of ECG Signals using SPIHT Algorithm", International Journal of Information & Communication Engineering, pp.219-225, 2005
- [7] Tahere Izack Mohammadpour, Mohammad Reza Karami Mollaei, "ECG compression with Thresholding of 2D -Wavelet transform Coefficients & run length coding," European Journal of Scientific Research, ISSN 1450, Vol. 27, No. 2, pp. 248-257, 2009
- [8] Santa Ktata, Kais Ouni, Nouredine Ellouze, "A Novel Compression Algorithm for Electrocardiogram Signals based on Wavelet Transform and SPIHT," World Academy of Science Engineering & Technology, pp. 855-860, 2009.
- [9] Morteza Moazami-Goudarzi, and Mohammad. H. Moradi, "Electrocardiogram Signal Compression using Multiwavelet Transform", Transactions on Engineering, Computing and Technology Vol. 6, pp.332-336, June 2005.
- [10] Sarada Prasad Dakua1 and Jyotinder Singh Sahambi, "Lossless ECG Compression for Event Recorder based on Burrows-Wheeler

- Transformation and Move-To-Front Coder”, International Journal of Recent Trends In Engineering, Vol. 1, No. 3, pp.120-123, May 2009.
- [11] Tatiparti Padma, M. Madhavi Latha, Abrar Ahmed, “ECG Compression and Labview implementation”, J. Biomedical Science and Engineering 2, pp.177-183, 2009
- [12] Vibha Aggarwal and Manjeet Singh Patterh, “Quality controlled ECG Compression using Discrete Cosine Transform (DCT) and Laplacian Pyramid (LP),” Multimedia, Signal Processing and Communication Technologies, pp .12-15, 2009.
- [13] Luisa F polania, Rapheal E Carrillo, Manuel Blanco-Velasco, Kenneth E Barner, “Exploiting prior knowledge in compressed sensing wireless ECG systems”, IEEE Journal of Biomedical and health Informatics, vol. 19, No. 2, pp 508-518, March 2015
- [14] Mohammed Abo-Zahhad, Sabah M Ahmed, Ahmed Zakaria, “An Efficient Technique for Compressing ECG Signals Using QRS Detection, Estimation, and 2D DWT Coefficients Thresholding”, Modelling and Simulation in Engineering, Hindawi Publishing Corporation Volume 2012 (2012), pp 1-10, 2012.
- [15] Sarita Mishra, Debasmit Das, Roshan Kumar, Parasuraman Sumathi, “ A power line interference canceler based on sliding DFT phase locking scheme for ECG signals”, IEEE transactions on Instrumentation and Measurement”, Vol. 64, No. 1, pp 132-142, January 2015
- [16] Yao Zau, Jun han, SizhongXuan, Shan Huang Xinqian Weng, Debin Fang, Xiaoyang Zeng, “An energy efficient design for ECG recording and R peak detection based on wavelet Transform”, IEEE transactions on circuits and systems, Vol. 62, No. 2,pp 119-123, Feb 2015
- [17] B Somanath Nair, “Digital Signal Processing”, PHI Edition, 2nd Edition, 2005.
- [18] Xuejing Kang, Feng Zhang, Ran Tao, “Multi channel random discrete fractional fourier transform”, IEEE Signal Processing letters, Vol 22, No. 9, pp 1340-1344, September 2015.
- [19] Hong Wen, Jie Tang, Jinsong WU, Huan Huan Song, Ting Yong WU, Bin WU, Pin-HO, Shi-Chao LV, Li-Min Sun, “A cross layer secure communication model based on Discrete Fractional Fourier Transform”, IEEE Transactions on emerging topics in computing”, Vol 3, No. 1, Mar 2015.
- [20] Surekha KS, BP Patil, “ECG signal compression using hybrid 1D and 2D Wavelet transform” IEEE Science and Information Conference (SAI) 2014, pp 468-472, 2014
- [21] <http://ecg.mit.edu/>

# Feature Selection Based on Minimum Overlap Probability (MOP) in Identifying Beef and Pork

Khoerul Anwar

Departement of Computer Science  
and Electronics  
Universitas Gadjah Mada  
Yogyakarta, Indonesia

Agus Harjoko

Departement of Computer Science  
and Electronics  
Universitas Gadjah Mada  
Yogyakarta, Indonesia

Suharto Suharto

Departement of Computer Science  
and Electronics  
Universitas Gadjah Mada  
Yogyakarta, Indonesia

**Abstract**—Feature selection is one of the most important techniques in image processing for classifying. In classifying beef and pork based on texture feature, feature overlaps are difficult issues. This paper proposed feature selection method by Minimum Overlap Probability (MOP) to get the best feature. The method was tested on two datasets of features of digital images of beef and pork which had similar textures and overlapping features. The selected features were used for data training and testing by Backpropagation Neural Network (BPNN). Data training process used single features and several selected feature combinations. The test result showed that BPNN managed to detect beef or pork images with 97.75% performance. From performance a conclusion was drawn that MOP method could be used to select the best features in feature selection for classifying/identifying two digital image objects with similar textures.

**Keywords**—overlap; feature selection; best feature; minimum overlap probability (MOP); identifying

## I. INTRODUCTION

Image identification based on texture features from many images with varying feature types,  $X = (x_i, i=1..M)$  is a difficult task. Multi-features of multi-objects for task of certain applications have three probabilities, i.e. relevant, irrelevant, and redundant features [1]. Similarly, according to [2-3], there are three main matters related to feature, i.e., 1) strong relevant, 2) weakly relevant and 3) irrelevant. Based on the argument, not having good knowledge on texture feature will be a problem in determining the best feature which can be used as a key for classification or identification. Therefore, feature selection process is important.

Feature selection is very effective in supporting performance in special tasks [3-7]. Several special tasks in image and computer vision processing are classification [8-10], clustering [11], computational neuroscience, imaging genomics [12 and 13], protein property prediction [14], text mining, image annotation, [15 and 16]. Feature has become an important part in the study of image and computer vision processing [17-18]. Feature is unique identity of an image. Unique identity of an image can be used as a key to recognize an image or can also be used to identify an image from another. Key feature is urgent when certain applications involve hundreds of data with tens of characteristics [1, 12 and 18]. In reality, a feature of two different objects for the same feature with different values is a problem on its own [18].

Feature selection is one of the main tasks in classification. In a large feature collection, it's possible that some or all of them are irrelevant or redundant features. Feature can be collected from extraction on three parts, i.e. texture, shape, and color of object. In this study, feature discussion is focused on texture feature. Feature separation must be performed with the correct extraction because the extraction feature would be used to train classifier. Selecting key feature is a part of a process to improve accuracy in classifier performance [19]. One of the feature selection techniques is selecting features with minimum redundancy criteria in the classification process [7]. Feature selection technique by minimum redundancy is also used by [4] for classification. The main function in selection in choosing minimum redundancy features from two objects which can be used to make distinct classes [2]. However, feature selection of both is used in multi-label features in single object. Unlike previous researchers [2],[4],[8] in this paper the writer performed multi-label feature selection for multi-object on identification of digital images of beef and pork. The selection criteria are features relevant with accuracy of classification. The writer suggested that a feature is relevant when it has minimum overlap. The challenge was choosing one or a few of the feature overlaps as candidates for the best feature. The basic assumption was the smaller the range of overlapping value of a feature, the better it was to select the feature to win the selection (key feature). The basic problem for feature overlap is the feature group isn't a comprehensive representation [20] of the value of a feature as a target because it still contains the value of other features.

In this paper, the writer offered a feature selection to obtain the best features from two groups of features from two digital images with similar textures by **MOP** method. The identity of each feature was determined by min-max values then calculating the overlapping value of each feature. Then overlap probability of each feature was calculated. The next step was selecting key (best) features by applying threshold on probability values. Features with probability less than the threshold won the selection.

This paper was arranged as follows: Section 2, describes relevant scientific works. Section 3, described methodology. Explanation on the MOP the author offered is reviewed in Section 4, Test and Result are described in Section 5 and Discussion in Section 6. Lastly is Conclusion.

## II. RELATED WORK

Kamyab and Eftekhari [21] in their paper discuss a special study on the usage of Multimodal Optimization (MO) method for feature selection. To do this, Evolutionary Algorithms (EAs) modification from several famous methods based on and a proposed *niching* method called GA\_SN\_CM are used for feature selection task and is compared with several famous EA-based methods for feature selection to study the strength of MO method on improving the result of feature selection.

Sotoca and Filiberto [22] in their paper on feature selection, or variable selection, select the most relevant features (attributes) of a group of variable data. In this framework, relevant term refers to the effect of given features or feature set to obtain possibility of minimum error in classification or recognition of classification problems.

Al-Ani, Alsukker and Rami [23] propose differential evolution algorithm for wrapper feature selection which uses the simplest yet effective way to narrow the search without removing any feature. A number of dataset with different sizes are used to evaluate the performance of the proposed method, which can give good indication on exploration and exploitation of the ability.

Kabir, Islam and Kazuyuki [24] in their paper proposes a new algorithm called constructive approach for Feature Selection (CAFS) based on wrapper approach concept consecutive search strategy. As a learning model, CAFS employs three layers of feed-forward Neural Network (NN). The proposed technique combines feature selection (FS) with NN architecture determination. It uses constructive approach which involves correlation information in selecting features and determining network architecture.

Lutu and Engelbrecht [25] in their paper discuss algorithm for feature selection in data mining prediction for classification problem by trying to categorize them to select relevant and not excessive features for classification task. A relevant feature is defined as one which correlates with target functions. As excessive feature is defined as one which correlates with other features. In this writing, they propose a new algorithm by combining the usage of certain threshold values and decision rules to select feature subset

Hanchuan, Fuhui and Chris [26] combine max-dependency, max-relevance and min-redundancy for feature selection. In reciprocal information, the purpose of feature selection is to discover a set of  $S$  features with  $m$  feature ( $x$ ), which together have the biggest dependency in target class  $c$ . This scheme is called max-dependency. But it's difficult to do. The alternative is max-relevance which is searching for features which fulfill max-dependency value by average value of all reciprocal information values between features of individual  $x$  and class.

## III. FRAME WORK MOP

Feature selection was performed to get the best features from feature set of beef and pork for classification task. Every feature set has 20 features where every feature consists of 200 data. The problem was the values of extraction features between beef and pork from the same feature name didn't

produce independent features, but feature overlaps instead. Fig 1 shows the framework for feature selection by MOP method. The focus on this stage is making a model to calculate area overlap of every feature between features of digital images of beef ( $F_s$ ) and pork ( $F_b$ ). The early stage is determining the area of each feature of  $F_s$  and  $F_b$ . It's continued by calculating the area overlap of both. Then, probability value of every feature was calculated. Lastly, overlap probability value was selected. The constructed model architecture is shown in Fig 1.

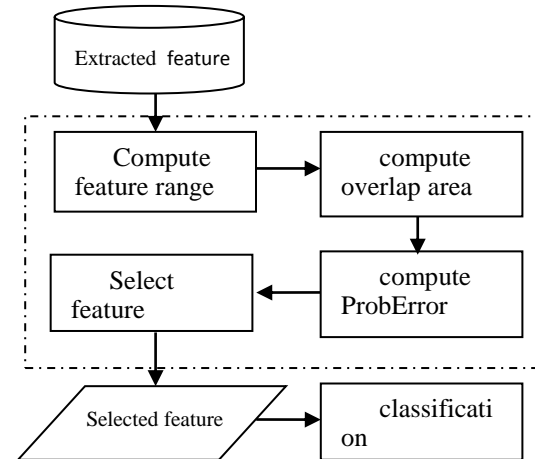


Fig. 1. Framework of feature selection by MOP

To determine whether selected features are the best features as expected, selected features were tested on artificial neural network (ANN).

### A. Extraction feature

Extraction is a pre-processing stage which is a basic stage to get maximum data before processing. Extraction was performed on each image to determine texture characteristics. The feature which became the object of the writer's research was the feature of the texture of digital images of beef and pork. The main thing to get was strong features which could be used to differentiate the texture of both. As usual, to get features in pre-process, the study conducted extraction process of both by several types of features which have been used by previous researchers. Some of those features were used to look for unique features from the extracted images. The extraction model in this study was gray level co-occurrence matrix (GLCM) method. GLCM is a tabulation of how often different combinations of gray level co-occurrence matrix are found in image section or images [27]. Calculation of texture feature used GLCM to get sizes of variations in intensity (i.e, image texture) in pixels which were focused on. Co-occurrence matrix was calculated by two parameters, which is relative distance between  $d$  of pixel pair measured in total pixel and their relative  $\theta$  orientation. These two parameters were expected to find special characteristics of two digital images of beef and pork. Unique features expected to be found maximally from the digital images were: autocorrelation, contrast, correlation, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, maximum probability, sum of square variance, sum average, sum variance, sum entropy, difference variance, difference entropy, information of correlation, Inverse difference normalized,



inverse difference moment normalized. The extraction result was numeric. The numbers were the data or sources of data processing. The extraction result showed the group of feature values for digital images of beef  $S = (x_1, x_2, x_3, \dots, x_n)$  and for digital images of pork,  $B = (x_1, x_2, x_3, \dots, x_n)$ , with  $x$  being feature names.  $x$  feature has a group of value from extraction of a number of extracted images. The names and formulas of extraction features used for classification/identification of beef and pork in this research was cited from [28] [29], as shown in Table I

TABLE I. EXTRACTION FEATURE

No	Features name	Formulation
1	Autocorrelation	$\sum_{i=0}^n \sum_{j=0}^n (i,j) P_{r,\theta}^t(i,j)$
2	Contrast	$\sum_{n=0}^{n_c-1} n^2 \left\{ \sum_{i=1}^{n_c} \sum_{j=1}^{n_j} p(i,j),  i-j  = n \right\}$
3	Correlation	$\frac{\sum_{i=0}^n \sum_{j=0}^n (i - \mu_{x,r,\theta}^t)(j - \mu_{y,r,\theta}^t) P_{r,\theta}^t(i,j)}{\sigma_{x,r,\theta}^t \sigma_{y,r,\theta}^t}$
4	Cluster Prominence	$\sum_{i=0}^n \sum_{j=0}^n ((i - \mu_{x,r,\theta}^t) + (j - \mu_{y,r,\theta}^t))^4 P_{r,\theta}^t(i,j)$
5	Cluster Shade (skewness)	$\sum_{i=0}^n \sum_{j=0}^n ((i - \mu_{x,r,\theta}^t) + (j - \mu_{y,r,\theta}^t))^3 P_{r,\theta}^t(i-j)$
6	Dissimilarity	$\sum_{i=0}^n \sum_{j=0}^n (i-j) P_{r,\theta}^t(i,j)$
7	Energy	$\sum_{i=0}^n \sum_{j=0}^n P_{r,\theta}^t(i,j)^2$
8	Entropy	$\sum_{i=0}^n \sum_{j=0}^n P_{r,\theta}^t(i,j)^2 \log_2(P_{r,\theta}^t(i,j))$
9	Homogeneity	$\sum_{i=0}^n \sum_{j=0}^n \frac{1}{1 + (i-j)^2} P_{r,\theta}^t(i,j)$
10	Maximum probability	$\frac{\text{argmax}}{i,j} P_{r,\theta}^t(i,j)$
11	Sum of squares: Variance	$\sum_i \sum_j (i - \mu)^2 \cdot p(i,j)$
12	Sum average	$\sum_{i=c}^{2N_c} i p_{x+y}(i)$
13	Sum variance	$\sum_{i=c}^{2N_c} (i - S_{ent})^2 P_{x+y}(i)$
14	Sum entropy	$-\sum_{i=c}^{2N_c} p_{x+y}(i) \log(P_{x+y}(i)) = S_{ent}$
15	Difference variance	$\sum_{i=0}^{N_c-1} i^2 p_{x-y}(i)$
16	Difference entropy	$-\sum_{i=c}^{N_c-1} p_{x-y}(i) \log(P_{x-y}(i))$
17	Information measure of correlation1	$\frac{HXY - HXY1}{\max\{HX, HY\}}$ <p>HX, HY adalah entropy pada px dan py, sehingga</p> $HXY = -\sum_{i=0}^n \sum_{j=0}^n P(i,j) \log(P(i,j))$ $HXY1 = -\sum_{i=0}^n \sum_{j=0}^n P(i,j) \log(P_x(i)P_y(j))$

No	Features name	Formulation
		$HXY2 = -\sum_{i=0}^n \sum_{j=0}^n P_x(i) P_y(j) \log(P_x(i)P_y(j))$
18	Informaiton measure of correlation2	$(1 - \exp -2(HXY2 - HXY) )^{1/2}$
19	Invers Difference Normalized	$\sum \frac{C(i,j)}{1 +  i-j }$ <p>Dimana</p> $C(i,j) = \frac{p(i,j)}{\sum_{i,j=1}^n p(i,j)}$
20	Invers Difference Moment Normalized	$\sum \frac{C(i,j)}{1 + (i-j)^2}$

**B. Feature Range**

Based on the extraction, the value of extraction features for 20 type of the features showed there was no particular feature which has independent range or categorized as strong relevance. Instead there were overlaps. So, in this study, the value of feature overlap received special attention or led to further study. This problem required certain formulation which can be used to determine the values of features in overlap area [16]. This method aims to get features with minimum overlap range probability and select features by certain thresholds to get the best features. Feature range is a range formed by minimum and maximum values. The formula was as follows:

1) *Max Value*: Maximum value is the highest value of a feature of a data set.

$$\text{Max}_x = \max(x_1 : x_m) \quad (1)$$

2) *Min Value*: Minimum value is the lowest value of a feature.

$$\text{Min}_x = \min(x_1 : x_m) \quad (2)$$

3) *Feature Range*: Area marked by minimum and maximum value limits of a feature

$$\text{Fitur}_x = (\min_x : \max_x) \quad (3)$$

$x_1 : x_m$  is a group of value of the 1<sup>st</sup> x feature to  $m$  amount of data

Ranges of beef features ( $F_s$ ) and pork features ( $F_b$ ) were areas formed by  $\text{Min}_x$  and  $\text{Max}_x$  (3) of extracted data. These min-max values were respectively used as the lower limits and upper limits of feature areas. So the area of every  $F_s$  and  $F_b$  could be determined. To determine overlapped area visually between  $F_s$  and  $F_b$  when interacting, each feature area was visualized in two dimensions (2D). The feature areas could be formed by giving range values of features to  $x$  and  $y$  axis. Therefore, value range in  $x$  axis was  $(\text{min}, \text{max})$ , the same value range applied to  $y$  axis. When similar features ( $X$ ) from beef and pork image data were described in the field, the features could be analyzed.

*Fb Range*, Ordinate point  $(x_1, y_1)$  was the lower left corner point or equaled to feature value  $x_1(\text{min}, \text{min})$ ,  $(x_2, y_1)$  was the lower right corner point which equaled to the feature value  $x_1(\text{max}, \text{min})$ ,  $(x_1, y_2)$  was the upper left corner point equaled to feature value  $x_1(\text{min}, \text{max})$  and  $(x_2, y_2)$  was the upper right

corner point equaled to feature value  $x_1(max,max)$ .  $Fb$  area could be determined based on the ordinates, so  $Fb$  was

$$Fb = ((x_1, y_1), (x_2, y_1); (x_1, y_2), (x_2, y_2)) \quad (4)$$

$Fs$  Range, Ordinate point  $(x_3, y_3)$  was the lower left corner point or equaled to feature value  $x_1(min,min)$ ,  $(x_4, y_3)$  was the lower right corner point which equaled to the feature value  $x_1(min,max)$ ,  $(x_3, y_4)$  was the upper left corner point equaled to feature value  $x_1(max,min)$  and  $(x_4, y_4)$  was the upper right corner point equaled to feature value  $x_1(max,max)$ .  $Fs$  area could be determined based on the ordinates, so  $Fs$  was

$$Fs = ((x_3, y_3), (x_4, y_3); (x_3, y_4), (x_4, y_4)) \quad (5)$$

### C. Overlap

Overlap between  $Fs$  happen  $Fb$  happened when they surpassed the value ranges of two or more features. In this study, feature had value range (3). Based on (3)  $Fs(X_1)$  and  $Fb(X_1)$  features overlapped when the maximum values of  $Fs(X_1)$  were bigger than the minimum values of  $Fb(X_1)$  and the minimum values of  $Fs(X_1)$  were less than the minimum values of  $Fb(X_1)$ . In theoretical discussion, overlapping set is called intersection. The formulation to get intersection value was

$$Fs \cap Fb = \{x | x \in Fs\} \cap \{x | x \in Fb\} \quad (6)$$

Equation (5) was effective to determine intersection element. However, this study was aimed to determine range, so this study modified (6) to find intersection value. The term intersection in this study was called overlap.

Feature overlap area ( $Fo$ ):  $Fo$  was overlapping area of  $Fb$  and  $Fs$  areas. This area was determined by the positions of ordinates  $Fb$  (4) and  $Fs$  (5).  $Fo$  area could be determined by ordinates  $Fs$  and  $Fb$ , so  $Fo$  area was

$$Fo = ((x_3, y_3), (x_2, y_3); (x_3, y_2), (x_2, y_2)) \quad (7)$$

Ordinate points in (7) were overlapped ordinate points which were, respectively,  $(x_3, y_3)$  which equaled to feature value  $Fs$  for  $(min,min)$ ,  $(x_2, y_3)$  was feature values  $Fs$  and  $Fb$  ( $Fs$  min,  $Fb$  max),  $(x_3, y_2)$  equaled to feature values  $Fs$  and  $Fb$  ( $Fb$  max,  $Fs$  min), and  $(x_2, y_2)$  equaled to feature value  $Fs$  for  $(max,max)$ . Fig 2 shows overlap area between a feature of  $Fs$  and  $Fs$ .

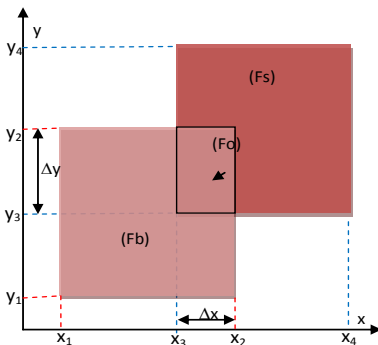


Fig. 2. Overlap area between  $Fb$  and  $Fs$

Fig 2 is interaction of  $Fs$  and  $Fb$  which shows an overlap between  $Fs$  and  $Fb$ .

### D. MOP

Probability of an incidence is a number which shows the possibility of an event. In this study, there was possibility of similar value between  $Fs$  and  $Fb$  in overlap area. The problem was how big the overlap was between them. Using set theory, the number of  $Fs$  members could be written as  $nFs$ , and the number of  $Fb$  set members could be written as  $nFb$ . Based on this, *probability of overlap area* (ProbArea) was defined as

$$ProbArea = \frac{|nFs \cap nFb|}{|nFs \cup nFb|} = \frac{|nFs \cap nFb|}{|nFs + nFb| - |nFs \cap nFb|} \quad (8)$$

In this study, the author modified (8) for the number of members of sets with feature area size. The formulation of each size is defined below:

*Size of feature areas of beef (Ls) and pork (Lb)*: The general formula of area size is length multiplied with width. Length in this case the length of  $Fs$  was the range of  $Fs$  along  $X$  axis, i.e. delta ( $\Delta x_s$ ) the distance between maximum point ( $x_4$ ) and minimum point ( $x_3$ ). While width of  $Fs$  was the range of  $Fs$  along  $Y$  axis, i.e. delta ( $\Delta y_s$ ) the distance between maximum point ( $y_4$ ) and minimum point ( $y_3$ ). While the length of  $Fb$  was the range of  $Fb$  along  $X$  axis, i.e. delta ( $\Delta x_b$ ) the distance between maximum point ( $x_2$ ) and minimum point ( $x_1$ ). While the width of  $Fb$  was the range of  $Fb$  along  $Y$  axis, i.e. delta ( $\Delta y_b$ ) the distance between maximum point ( $y_2$ ) and minimum point ( $y_1$ ). So,  $Ls$  was defined as:

$$\Delta x_s = |x_4 - x_3|$$

$$\Delta y_s = |y_4 - y_3|$$

$$\Delta x_s = \Delta y_s \quad (9)$$

where

$$Ls = |x_4 - x_3| * |y_4 - y_3| \quad (10)$$

Based on (9) and (10), so

$$Ls = \Delta x_s^2 = \Delta y_s^2 \quad (11)$$

and  $Lb$  could be defined as:

$$\Delta x_b = |x_2 - x_1|$$

$$\Delta y_b = |y_2 - y_1|$$

$$\Delta x_b = \Delta y_b \quad (12)$$

where

$$Lb = |x_2 - x_1| * |y_2 - y_1| \quad (13)$$

Base on (12) and (13), so

$$Lb = \Delta x_b^2 = \Delta y_b^2 \quad (14)$$

1) *Size of overlap area (Lo)*: To determine the length and width of  $Lo$  area, Fig 2 shows the length of  $Lo$  is  $\Delta x$  and the width  $\Delta y$ . Each could be calculated by the following equations:

$$\Delta x = |x_2 - x_3| \quad (15)$$

$$\Delta y = |y_2 - y_3| \quad (16)$$

Equations (15) and (16), for  $\Delta x = |x_2 - x_3|$ , point  $x_2$  was maximum value for  $Fb$  and point  $x_3$  was minimum value of  $Fs$ .  $\Delta y = |y_2 - y_3|$ , point  $y_2$   $x_2$  was maximum value for  $Fb$  and point  $y_3$  was minimum value of  $Fs$ . So based on (15) and (16) the size of overlap area ( $Lo$ ) was

$$Lo = \Delta x * \Delta y \quad (17)$$

Based on (11) and (14),  $L_s$  and  $L_b$  were areas with the same length on both sides. So  $Lo$  in (17), could be written as

$$Lo = \Delta x^2 \quad (18)$$

Using  $Lo$  (18) for feature values in this area was indicated to cause problem for the process of identifying images of beef or pork. The problem happened in the area maybe due to duplication of values of features of beef and pork. Thus, the bigger the value of  $Lo$ , the bigger the amount of duplication of members of features of beef and pork, and vice versa. It should be noted that this area was formed by the range of feature values, so the overlap area of every feature wasn't absolute at certain amounts because the range was influenced by the area stability of each feature. However, by using area range based on min-max of the features, the system was still able to get overlap area. The possibility of overlap or overlap probability was the main focus of this study. To determine overlap probability of every feature between  $F_s$  and  $F_b$ ,  $Lo$  could be compared with the size of all features ( $F_s + F_b$ ). In this study, computation of overlap probability by the author was called error probability ( $ProbError$ ). The formula was

$$ProbError = 2 * Lo / (L_s + L_b) \quad (19)$$

Equation (19) meant that the smaller the value of  $Lo$ , the smaller the value of  $ProbError$ . Conversely, the bigger the value of  $Lo$ , the bigger the value of  $ProbError$ .

The author used this  $ProbError$  value as data to select the best features. Selection was performed by giving threshold value  $< 10\%$ . The author named this method *Minimum Overlap Probability (MOP)* method.

#### E. MOP Algorithm:

1) Algorithm of feature selection by *Minimum Overlap Probability (MOP)* method

a) Calculating minmax values of extraction features of digital images of beef and pork

b) Determining feature area ( $F_x$ ) of digital images of beef and pork.

- If  $F_x$  had no overlap (independent) it's a selected feature
- If  $F_x$  was a subset or superset of each other,  $F_x$  wasn't a selected feature (rejected)
- If the calculation delta  $F_x$  of a group of feature database wasn't in a or b process, the process was continued (process 3).

c) Calculating  $ProbError$  value

d) Determining threshold (as filter of selection of selected features)

e) Finding features with  $ProbError$  less than threshold

f) Selected features

#### F. MOP Flowchart

Algorithm of feature selection by MOP method is illustrated in the flowchart in Fig 3.

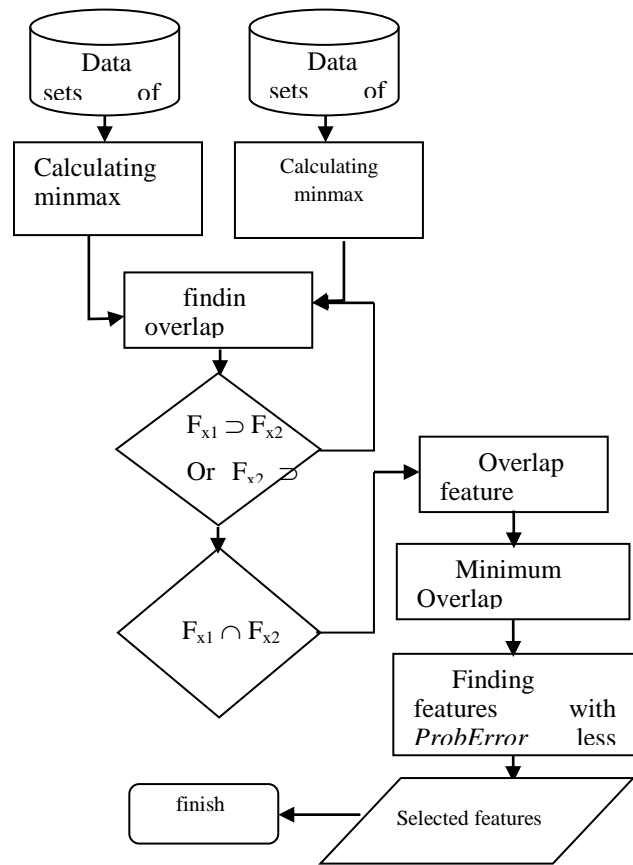


Fig. 3. Flowchart of Feature selection by MOP

#### G. Testing the selected features

Testing the features selected conducted on artificial neural networks. This testing is done to determine the effect on the accuracy of results. Type of neural network used is a multi-layer back propagation neural network.. The network architecture used here was I-H-O i.e. input layer, hidden layer and lastly layer output. To determine the correlations of selected features and accuracy of network classification, input layer was set up for several nodes. Meanwhile, output layer was set up was two nodes. In training stage, the target classes were label 00 for pork, 11 for beef. To support the performance of the network, the selected learning method was levenberg marguad. It's because this method has the best accuracy compared with other learning methods.

### IV. EXPERIMENT AND RESULT

#### A. Experiment

Image data was acquired by mobile digital camera at five mega pixel (5MP). Total data was 400 images with 200 images each for beef and pork. The dimension of digital images was 255 x 255 and in JPEG format. Data was pre-processed by converting RGB to gray, filtering images by gabor. Total and names of extraction features are written in Table I. Threshold value was 10%. The process in this experiment is as follows:

- 1) Reading Fb and Fb data sets
- 2) Determining the range of min and max values of Fs and Fb (3)
- 3) Calculating proberror value by equation (14).
- 4) *Selecting features.* In this selection process, a criterion was used to select the best features. The best features met the following criterion:

$$f(x) = \begin{cases} 1, & \text{Proberror} < \text{threshold} \\ 0, & \text{Proberror} \geq \text{threshold} \end{cases} \quad (15)$$

Note (15) f(x) is selected feature, threshold < 10%. Criterion f(x) =1 means fulfilling requirement or accepted, while f(x)=0 means not fulfilling the criterion or rejected.

- 5) The final step was testing selected features on neural network.

### B. Result

The range values of Fs and Fb, and error probability of every feature from the implementation of MOP method produced the result shown in Table II

TABLE II. RANGE VALUES AND ERROR PROBABILITY

Feature Name	Fs Range		Fb Range		Prob Error
Maximum probability	0.61	0.76	0.74	0.95	0.06
Contrast	0.12	0.44	0.02	0.15	0.07
Difference variance	0.12	0.44	0.02	0.15	0.07
Sum average	2.51	3.10	2.13	2.58	0.07
Autocorrelation	1.95	3.32	1.35	2.10	0.08
Energy	0.39	0.60	0.56	0.90	0.08
Sum entropy	0.85	1.37	0.25	0.94	0.08
Entropy	0.92	1.67	0.26	1.05	0.09
Dissimilarity	0.11	0.32	0.02	0.14	0.10
Sum of squares: Variance	1.92	3.49	1.33	2.14	0.10
Sum variance	4.28	7.07	4.03	4.59	0.10
Roughness	0.05	0.12	0.03	0.06	0.11
Difference entropy	0.34	0.72	0.10	0.41	0.11
Informaiton measure of correlation2	0.63	0.74	0.44	0.67	0.13
Homogeneity	0.85	0.95	0.93	0.99	0.14
Cluster Shade (skewness )	11.43	23.66	10.67	13.29	0.14
Homogeneity	0.86	0.95	0.93	0.99	0.15
Inverse difference normalized (INN)	0.97	0.99	0.98	1.00	0.33
Correlation	0.80	0.90	0.84	0.96	0.38
Correlation	0.80	0.90	0.84	0.96	0.38
Information measure of correlation1	0.32	0.72	0.09	0.42	0.16
Regularity	0.00	0.03	0.01	0.04	0.69
Inverse difference moment normalized	0.99	1.00	1.00	1.00	100.00
Cluster Prominence	135.56	230.61	135.54	170.01	Error

On Table II, feature selection was performed by using the determined filter value (threshold < 10%). The result of the selection was names of selected features shown in Table III.

TABLE III. SELECTED FEATURES

No	Feature name	Fs Range		Fb Range		Prob Error (%)
1	Maximum probability	0.61	0.76	0.74	0.95	6
2	Contrast	0.12	0.44	0.02	0.15	7
3	Difference variance	0.12	0.44	0.02	0.15	7
4	Sum average	2.51	3.10	2.13	2.58	7
5	Autocorrelation	1.95	3.32	1.35	2.10	8
6	Energy	0.39	0.60	0.56	0.90	8
7	Sum entropy	0.85	1.37	0.25	0.94	8
8	Entropy	0.92	1.67	0.26	1.05	9

Table III shows selection result with error probability values of 6%,7%,8% and 9%. Some features had the same error probability values, e.g. contrast, difference variance, sum average with error probability value of 7%, autocorrelation, energy, sum entropy with error probability value of 8%.

The test on selected features for classification task was performed on neural network. The architecture of neural network was five nodes of input, five nodes of hidden layer and 2 nodes of output (5I-5H-2O). There were two models of the test, first using the best feature as single input and second using combination of some of the best features. The performance of neural network for data training by best feature input (maximum probability) produced 95.50%. A different result was shown by several combinations of the best features which produced 100%. The result of data testing by some combinations of features as input of neural network was shown in Table IV.

TABLE IV. ACCURACY OF FEATURE COMBINATION ON NN

Combinations of features	accuracy (%)
1,2	96.75
1,2,3	96.00
1,2,3,4	95.00
1,2,6,8	97.75
1,2,3,4,5	92.75
1,2,3,4,5,6	95.50
1,2,3,4,5,6,7	97.50
1,2,3,4,5,6,7,8	94.75

Table IV is the result of data testing based on data classification. Combination of features 1,2,6,8 had the highest accuracy of neural network (97.75%), while combination of features 1,2,3,4,5 had the lowest accuracy (92.75%). It showed that classification by combination of selected features produced accuracy of performance of neural network above 92.00%

### V. DISCUSSION

Feature selection by MOP with threshold 0.1 selected maximum probability, contrast, Difference variance, Sum average, Autocorrelation, energy, Sum entropy, and entropy as the best features from 20 feature candidates. It meant that

these features in digital images of beef and pork had smaller overlap values than other features. Combinations of selected features were used to train network and then testing was performed using new data, showing the best features could support network performance. The lowest network accuracy was in feature combination 1,2,3,4 and 5 with 92.75% accuracy or error level of 7.25%. The best accuracy was in feature combination 1,2,6 and 8 which had network performance with 97.75% accuracy or error level of 2.25%. It showed that feature combinations influenced accuracy of classification. Based on the result of the test, the selected features were correct and could be used as unique characteristics to identify beef or pork by digital image.

## VI. CONCLUSION

Overlap probability can be used to select the best features of some of the features that have value overlap one another. MOP method could be used as one of the solutions for selecting the best or strongest features of two objects with feature overlap.

The selected feature is a maximum probability, contrast, energy and entropy is the best feature based on the results of testing with artificial neural networks. It is derived from the performance of the neural network with an accuracy rate of 97.75%. In other performance error rate of 2.25%.

Future work will be the development Minimum Overlap Probability method to determine the correlation between the selected feature.

## ACKNOWLEDGMENT

The writer pays tribute to: (1) Ministry of Research and Technology and Higher Education, Republic of Indonesia, for the funding source (BPDN); (2) STMIK Pradnya Paramita; Malang ([www.Pradnya-paramita.ac.id](http://www.Pradnya-paramita.ac.id)); (3) Department of Computer Science and Electronics, Universitas Gadjah Mada (<http://mkom.ugm.ac.id>) Yogyakarta Indonesia.

## REFERENCES

- [1] B. Xue, M. Zhang, and W. N. Browne, "Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach." *IEEE Transactions On Cybernetics*, pp 1656-1671, 2013.
- [2] J.R. Vergara, and P. A. Estevez, "A review of feature selection methods based on mutual information." *Neural Comput & Applic* xxiv, pp 175-186, 2014.
- [3] L.Huawen, W. Xindong, and Z. Shichao, "A new supervised feature selection method for pattern classification." *Computational Intelligence* (Blackwell Publishing Inc) XXX, No. 2, pp 342-361, 2014.
- [4] L.Yaojin, H. Qinghua, L.Jinghua, and D. Jie, "Multi-label feature selection based on max-dependency and min-redundancy." *Neurocomputing*, Elsevier, pp 92-103, 2015.
- [5] H. Zuhair, A.Selmat, and M.Salleh, "The Effect of Feature Selection on Phish Website Detection." (IJACSA) *International Journal of Advance Computer Science an Applications* VI, No. 10, pp 221-232, 2015.
- [6] S. Elseuofi, and W. Awad, "Toward Accurate Feature Selection Based on BSS-GRF." (IJACSA) *Intenational Journal of Advance Computer Science an Applications* V, No. 8, pp 55-59, 2015.
- [7] P. Hanchuan, L. Fuhui, and D. Chris, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy." *IEEE Transactions On Pattern Analysis And Machine Intelligence*, IEEE, Vol 27, No. 8 pp 1226-1238, August 2005.
- [8] C. Bolun, C. Ling, and C. Yixin, "Efficient ant colony optimization for image feature selection." *Signal Processing*, pp 1566-1576, 2013.
- [9] L. Xiaoming, and T. Jinshan, "Mass Classification in Mammograms Using Selected Geometry and Texture Features, and a New SVM-Based Feature Selection Method." *Systems Journal, IEEE*, Vol. 8, No. 3, pp 910-920, September 2014.
- [10] Y. V. Goncharov, I. B. Muchnik, and L. V. Shvartser, "Feature Selection Algorithm in Classification Learning Using Support Vector Machines." *Computational Mathematics and Mathematical Physics*, pp 1243-1260, 2008.
- [11] A. Emre, and K. Adem, "Simultaneous feature selection and ant colony clustering." *Procedia Computer Science*, Elsevier, pp 1432-1438, 2011.
- [12] W.Hua, et al. "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort." *Bioinformatics*, Vol 28, No. 2, pp 229-237, 2012.
- [13] W. Awada, T.M. Khoshgoftaa, D.Dittman, R. Wald, and A. Napolitano, "A Review of the Stability of Feature Selection Techniques for Bioinformatics Data." *IEEE*, pp 356-363, 2012.
- [14] W. De, Y. Liang, F. Zhengqi, and X. Jingbo, "Prediction of Thermophilic Protein with Pseudo Amino Acid Composition: An Approach from Combined Feature Selection and Reduction." *Protein And Peptide Letters*, 684-689, March 2011.
- [15] D. D Lewis, "Feature selection and feature extraction for text." *Proceeding. Workshop Speech Natural Language*. Stroudsburg, pp 212-217, 1992.
- [16] D. Wang, F.Nie, and H. Huang, "Feature Selection via Global Redundancy Minimization." *IEEE Transactions On Knowledge And Data Engineering*, Vol 27, No. 10, pp 2743-2755, October 2015.
- [17] S. An, and X. Fan, "Study on Method of Feature Selection in Speech Content Classification." (IJACSA) *International Journal of Advanced Computer Science and Applications*, Vol 54, No. 4 pp 71-75, 2014.
- [18] S.N Devi, and S.P Rajagopalan. "A study on Feature Selection Techniques in Bio-Informatics." (IJACSA) *International Journal of Advance Computer Science dan Applications*, Vol 2, No. 1, pp 138-144, 2011.
- [19] J. Liu, and G. Wang, "A Hybrid Feature Selection Method for Data Sets of thousands of Variables." *IEEE*, pp 288-291, 2010.
- [20] N. Valliammal, and S.N. Geethalakshmi, "An Optimal Feature Subset Selection for Leaf Analysis." *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, Vol 6, No. 2, pp 191-196, 2012.
- [21] C. Ding, and H. Peng, "Minimum Redundancy Feature Selection From microarray Gene Expression Data." *Journal of Bioinformatics and Computational Biology*, pp 185-205, 2005.
- [22] Kamyab, Shima, and Mahdi Eftekhari. "Feature selection using multimodal optimization techniques." *Neurocomputing*, pp 586-597, 2016.
- [23] J.M. Sotoca, and P. Filiberto, "Supervised feature selection by clustering using conditional mutual information-based distances." *Pattern Recognition*, Elsevier, Vol 43, pp 2068-2081, 2010.
- [24] A. Al-Ani, A. Alsukker, and N. K. Rami, "Feature subset selection using differential evolution and a wheel based search strategy." *Swarm and Evolutionary Computation*, Elsevier, pp 15-26, (2013).
- [25] M.M Kabir, M.MIslam, and M. Kazuyuki, "A new wrapper feature selection approach using neural network." *Neurocomputing*, Elsevier, vol 73, pp 3273-3283, 2010.
- [26] P.E. Lutu, and A.P. Engelbrecht, "A decision rule-based method for feature selection in predictive data mining." *Expert Systems with Applications*, Elsevier, Vol 47, pp 602-609, 2010.
- [27] H. Shahera, and S.Seiichi, "Features for Texture Analysis." *SICE Annual Conference*. Akita, Japan: SICE, pp 1939-1744, 2013.
- [28] L.K Soh, and C. Tsatsoulis. "Texture Analysis of SAR Sea Ice Imagery Using Gray Level Co-Occurrence Matrices." *Transactions on Geoscience and Remote Sensing, IEEE*, Vol 37 pp 780-795, March 1999.
- [29] D.A. Clausi, and Y. Zhao, "Grey level co-occurrence integrated algorithm (GLCIA): a superior computational method to rapidly determine co-occurrence probability texture features." *Computer & Geosciences*, pp 837-850, 2003.

# Resource Allocation in Cloud Computing Using Imperialist Competitive Algorithm with Reliability Approach

Maryam Fayazi

Department of Computer, Ahvaz  
Branch, Islamic Azad University  
Department of Computer,  
Khouzestan Science and Research  
Branch, Islamic Azad University  
Ahvaz, Iran

MohammadReza Noorimehr

Department of Computer, Ahvaz  
Branch, Islamic Azad University  
Department of Computer,  
Khouzestan Science and Research  
Branch, Islamic Azad University  
Ahvaz, Iran

Sayed Enayatollah Alavi

Department of Computer  
Engineering, Faculty Engineering,  
Shahid Chamran University of  
Ahvaz  
Ahvaz, Iran

**Abstract**—Cloud computing has become a universal trend now. So, for users, the reliability is an effective factor to use this technology. In addition, users prefer to implement and get their work done quickly. This paper takes into account these two parameters for resource allocation due to their importance. In this method, the Imperialist Competitive algorithm with the addition of a cross layer of cloud architecture to reliability evaluation is used. In this cross layer, initial reliability is considered for all the resources and the implementation of their tasks and due to the success or failure of implementation, reliability of resources is increased or reduced. Reliability and makespan are used as a cost function in ICA for resource allocation. Results show that the proposed method can search the problem space in a better manner and give a better performance when compared to other methods.

**Keywords**—Imperialist Competitive algorithm; Reliability; makespan; Cloud Computing

## I. INTRODUCTION

Research on cloud computing is growing rapidly. Cloud computing means developing and applying computer technology on the Internet. Cloud computing is a network for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction[1]. In cloud computing, the end users have unlimited access to the resources and only pay for the resources they consume. Cloud computing basics is that the user data is not stored locally, but is stored in the internet data center. Companies that provide cloud computing services can manage and maintain the data centers [2, 3].

To provide services to end users, the cloud computing environment needs to be reliable and also well managed in such a way that it gives throughput in the lowest time. So, reliability and task scheduling are two important parameters. In this paper, considering reliability, the allocation of makespan resources was done using Imperialist Competitive algorithm and a cross layer in cloud architecture.

This paper is organized as follows. Section 2 explains some related works. Section 3 is dedicated to problem description and some meta-heuristic algorithms related to problems such as genetic algorithm and imperialist Competitive algorithm. Section 4 is a discussion about proposed approach that uses imperialist Competitive algorithm in it. Section 5 shows simulation evaluations and comparison to genetic algorithm. Section 6 is the conclusion.

## II. RELATED WORK

Many heuristic and meta-heuristic methods are prevented by different researchers for scheduling and resource allocation to tasks in cloud. The heuristic approach uses the concept of prioritization to schedule tasks. Algorithms such as max-min and min-min are part of the heuristic-based approach that divide scheduling to two steps: prioritizing the task and source selection. In priority task, each task is assigned a rating based on its priorities. In step source selection, a high-priority task is selected and is scheduled on the optimized resources that have completed the previous tasks. The meta-heuristic approach includes scheduling algorithms based on repeatable method to find the optimal solution. They provide an efficient way of moving quickly towards a very good solution. Many meta-heuristic approaches have been applied for solving workflow scheduling problems, including Genetic Algorithms and Imperialist Competitive Algorithm (ICA) [4].

One of important problems in cloud computing is makespan constraint. Most scheduling algorithms focus on makespan. In [5]-[10], Genetic Algorithm is used to reduce makespan of tasks allocation to resources. Yue Miao [11] by using firefly algorithm based on chaos algorithm tried to reduce the average time spent by subtasks in processing request tasks, and thus improve the efficiency of task processing and achieve a rational allocation of resources. Mizan et al. [12] proposed a modified task scheduling algorithm based on the concept of Bees life algorithm and greedy algorithm to gain optimistic value of service in hybrid cloud. The main idea of the system is to achieve an affirmative response from the end users and utilize the resources in a very transient manner.



Seidgar et al. [13] describe an approach incorporating simulation with imperialist competitive algorithm for the scheduling purpose having machine breakdowns and preventive maintenance activities. The objective is to minimize the makespan.

In addition of makespan, reliability function is a very important function for users [14]-[20].

Zhao et al. [14] proposed a DRR (Deadline, Reliability, Resource-aware) scheduling algorithm, which schedules the tasks such that all the jobs can be completed before the deadline, ensuring the Reliability and minimization of resources. Gartner [15] used a formal approach to define important terms like fault, fault tolerance, and redundancy. This leads to four distinct forms of fault tolerance and two main phases in achieving them: detection and correction. It shows that this can help to reveal inherently fundamental structures that contribute to understanding and unifying methods and terminology. By doing this, it surveys many existing methodologies and discuss their relations. The underlying system model is the close-to-reality asynchronous message-passing model of distributed computing. Zhang et al. [16] present BFT Cloud (Byzantine Fault Tolerant Cloud), a Byzantine fault tolerance framework for building robust systems involuntary-resource cloud environments. BFT Cloud guarantees robustness of systems when up to  $f$  of totally  $3f+1$  resource providers are faulty, including crash faults, arbitrary behaviors faults, etc. BFT Cloud is evaluated in a large-scale real-world experiment which consists of 257 voluntary-resource providers located in 26 countries. The experimental results shows that BFT Cloud guarantees high reliability of systems built on the top of voluntary-resource cloud infrastructure and ensures good performance of these systems. In [17], the different techniques of fault tolerance are presented. The main focus is on types of faults occurring in the system, fault detection, and recovery techniques. In [18], the existing fault tolerance techniques in cloud computing are discussed based on their policies, tools used, and research challenges. Cloud virtualized system architecture has been proposed. In the proposed system, autonomic fault tolerance has been implemented. The experimental results demonstrate that the proposed system can deal with various software faults for server applications in a cloud virtualized environment. Jhavar et al. [19] introduced an innovative, system-level, modular perspective on creating and managing fault tolerance in Clouds. They proposed a comprehensive high-level approach to shading the implementation details of the fault tolerance techniques to application developers and users by means of a dedicated service layer. In particular, the service layer allows the user to specify and apply the desired level of fault tolerance, and does not require knowledge about the fault tolerance techniques that are available in the envisioned Cloud and their implementations. In [20], a fault tolerance model for cloud computing is given and the Paper describes a model for Fault Tolerance in Cloud computing (FTMC). FTMC model tolerates the faults on the basis of reliability of each computing node. A Computing node is selected for computation on the basis of its reliability and can be removed, if it does not perform well for applications.

To increase the Quality of Service in cloud, it is necessary to implement a scheduling algorithm that in addition to makespan, considers the reliability of cloud resources during resource allocation. Recently in [21], makespan and reliability were discussed. A genetic algorithm has been proposed that schedules workflow applications in unreliable cloud environment and meets user defined QoS constraints. A budget constrained time minimization genetic algorithm has been proposed which reduces the failure rate and makespan of workflow applications. It allocates those resources to workflow application which are reliable. So, here with this motivation, work is done that reduces makespan with Imperialist Competitive algorithm to provide reliable machines for implementation of tasks.

### III. BACKGROUND

#### A. Problem Description

Earlier, the Cloud data center was composed of thousands of servers and hundreds of switches connecting the servers. Each server can host for tens of virtual machines [22]. From scheduling algorithms, it is expected to find a plan for each task in set  $T$  with desired quality of service constraints for users. The schedule should be one that reduces the failure rate and makespan.

#### B. Imperialist Competitive Algorithm

Imperialist Competitive Algorithm is defined as the optimization strategy based on social and political evolution of humans. The Main Basics of this algorithm are Assimilation, Imperialistic Competitive, and Revolution. For more accurate algorithm inspired social and political phenomenon of colonialism [23]. The Pseudo code for the algorithm is as follows:

- 1) Select some random points and initialize the empires.
- 2) Move the colonies toward their relevant imperialist (Assimilating).
- 3) If there is a colony in an empire which has lower cost than that of imperialist, exchange the Positions of that imperialist and the colony.
- 4) Compute the total cost of an empire (Related to the power of both imperialist and its colonies).
- 5) Pick the weakest colony from the weakest empire and give it to the empire that has the most likelihood to possess it (Imperialistic Competitive).
- 6) Eliminate the powerless empires.
- 7) If there is just one empire, stop if not go to 2.

In step 1, randomly some points are selected and some Empires are formed. Then, according to equations (1), (2), and considering N.C. in equation (3), for each empire, the same number of initial colonial countries are selected, randomly and given to  $n^{\text{th}}$  Empire.

$$C_n = \max_i \{c_i\} - c_n \quad (1)$$

$C_n$  :  $N^{\text{th}}$  Empire normalized cost

$c_n$  :  $N^{\text{th}}$  Empire cost

$\max_i \{c_i\}$  : Max cost between Empires

$$P_n = \left| \frac{c_n}{\sum_{i=1}^{N_{imp}} c_i} \right| \quad (2)$$

$P_n$  :  $N^{\text{th}}$  Empire normalized power

$N_{imp}$  : Number of Empires

$$N.C \text{ (Number of colonies of each Empire)} = \text{round} \left( \frac{P_n * N_{col}}{N_{col}} \right) \quad (3)$$

$N_{col}$  : total number of colonies

In step 2, Empires with the follow-up policy of assimilation efforts are selected to attract their colonies. In line with this policy, the colonial country moves the size of  $x$  units to the Empire and get a new position.  $X$  is a random number with uniform distribution (or any other suitable distribution).

In step 3, it is possible that the colony moves to the empire get better position than the empire (lower cost). In this case, empire and colony exchange and the algorithm continues with the empire country in a new position.

In step 4, according to formula 4, the power of an empire is equal to the central government plus a small percentage of the power of its colonies.

$$T.C_n = \text{Cost}(\text{empire}_n) + \xi \text{ mean} \{ \text{Cost}(\text{colonies of empire}_n) \} \quad (4)$$

In equation (4),  $T.C_n$  is the total cost of  $n^{\text{th}}$  Empire and  $\xi$  is a positive number that is considered usually between zero and one, and close to zero. Little consideration of  $\xi$ , makes the total cost of an empire, almost equal to the cost of the central government (the empire) and also increasing  $\xi$  increases the effect of the colonies cost of that empire in determining the total cost.

The total cost of empire is obtained by the following equation (5):

$$N.T.C_n = \max_i \{ T.C_i \} - T.C_n. \quad (5)$$

In equation (5),  $T.C_n$  is total cost of  $n^{\text{th}}$  empire and  $N.T.C_n$  is total normalized cost of that empire. Each empire with less  $T.C_n$ , will be more  $N.T.C_n$ . In fact, the empire with min cost has max power.

In step 5, there is a competition between empires on taking over the weakest colony of the weakest empire.

By using the total normalized cost, the possibility (power) of takeover of Competitive colony by each empire, is calculated as in equation (6):

$$P_{pn} = \left| \frac{N.T.C_n}{\sum_{i=1}^{N_{imp}} N.T.C_i} \right| \quad (6)$$

With the possibility of taking of the empire, for dividing these colonies between empires randomly but with probability of related to possibility of taking of empire; vector  $p$  is formed from the above probability values (equation (7)):

$$\mathbf{P} = \left[ P_{p_1}, P_{p_2}, P_{p_3}, \dots, P_{p_{N_{imp}}} \right] \quad (7)$$

Vector  $p$  in equation (7) has a size of  $1 * N_{imp}$  and is made from possibility takeover empires values. Then, the random vector  $R$ , is formed in the same size of vector  $p$ . arrays of this vector are random numbers with uniform distribution in the interval  $[0, 1]$ . Then, vector  $D$  is made from the following equation (8):

$$\begin{aligned} \mathbf{D} &= \mathbf{P} - \mathbf{R} = \left[ D_1, D_2, D_3, \dots, D_{N_{imp}} \right] \\ &= \left[ P_{p_1} - r_1, P_{p_2} - r_2, P_{p_3} - r_3, \dots, P_{p_{N_{imp}}} - r_{N_{imp}} \right] \end{aligned} \quad (8)$$

By using vector  $D$ , those colonies where the index in vector  $D$  is larger than others are given to the empire. In step 6, the empire that doesn't have any colony, is eliminated to be a colony.

#### IV. PROPOSED METHOD

This section discusses the proposed approach that uses Imperialist Competitive algorithm and cross layer for evaluating the reliability of virtual machines.

##### A. Evaluation Reliability Algorithm

In [20] a method is used to increase fault tolerance and this method is applied as cross layer in cloud architecture to evaluate the reliability of virtual machines and then the VM with a high reliability is selected. Reliability evaluation algorithm is applied on each VM. At first, each VM reliability will be set to 1. There is an adjustment factor  $N$  that controls reliability evaluation.  $N$  value is always greater than 0. The algorithm gets the RF factor from input. RF is a reliability factor that increases or decreases resource reliability. RF value depends on user decision. Reliability should be between 0, 1. If task implementation by resource is done before defined deadline, it is successful and it's reliability will be increased. Otherwise it is a failure and it's reliability will be reduced. Pseudo-code of reliability evaluation algorithm is shown in Figure 1. Table I gives a sample of VM reliability evaluation.

```

1 Start
2 Initialized Reliability: =1, N: =1;
3 Input RF;
4 Input processing node status;
5 If processing node Status =Pass then
6 Reliability: = Reliability + (Reliability * RF);
7 If N > 1 then
8 N: = N-1;
9 If Reliability > 1 then
10 Reliability: = 1;
11 Else if processing node Status = Fail then
12 Reliability: = Reliability - (Reliability * RF * N);
13 N: = N+1;
14 If Reliability < 0 then
15 Reliability: = 0;
16 End
    
```

Fig. 1. Pseudo-code of reliability evaluation algorithm

TABLE I. SAMPLE OF VM RELIABILITY EVALUATION

Time	VM1	Reliability
1	Fail	0.95
2	Fail	0.9
3	Success	0.945
4	Success	0.99
5	Fail	0.94

**B. Imperialist Competitive algorithm**

To achieve the best allocation of resources in cloud, the Imperialist Competitive algorithm as follows:

1) *Problem modeling*: in this method, n different tasks and m virtual machines are considered and also  $T_i$  means  $i^{th}$  task and  $VM_j$  means  $j^{th}$  virtual machines. It is done by using identification number of each virtual machine and the task. In Figure 1, there are 10 tasks and each task has a unique identification number from 0 to 5 and there are 3 virtual machines each having unique identification number from 0 to 2. In Imperialist Competitive algorithm, each country represents a mapping of tasks to available resources on the issue of scheduling tasks. We are looking for the best possible sequence of tasks to allocate to resources. In ICA, position of each country represents a solution for a problem that shows mapping tasks on resources. Figure 2 shows sample of mapping tasks to resources by using identification number of tasks and virtual machines.

Tasks	T0	T1	T2	T3	T4	T5
Virtual Machines	VM1	VM2	VM0	VM2	VM0	VM1

Fig. 2. Countries as resource allocation

2) *Initial Population*: In initial population, countries are generated randomly. In this problem, each country is an array with  $2*N_t$  length that first row is tasks and second row is resources and The task t is selected from set T that has not been allocated to any virtual machine. Then a virtual machine is selected randomly from set VM. This process continues until countries are generated. The number of countries are generated in initial population according to the size of initial population. Then, number of initial empires and colonies is calculated from follow equations (9), (10):

$$\text{Initial empires} = \text{Countries} * n \tag{9}$$

$$\text{All Colonies} = \text{Countries} - \text{Initial empires} \tag{10}$$

In equation (10), n is a percentage of total number of countries that in the proposed algorithm is considered to be 0.1. In this paper, initial population size and number of empires are 100, 10, respectively.

3) *Evaluation*: in this step, best country is found. So, fitness value is calculated based on makespan and reliability. A country having low makespan and high reliability will have low cost value compared to other countries. In the proposed method, makespan must be minimized. Makespan and task execution time are calculated respectively in equations (11), (12).

$$\text{Makespan } M(I) = \sum_{i=0}^n T_{eti} \tag{11}$$

$$T_{et} = (\text{length of task}) / (\text{MIPS of virtual machine}) \tag{12}$$

In this method, in addition to reduced makespan, reliability also will increase. So, the solution giving a low makespan and high reliability will be selected. Reliability is obtained from cross layer.

So, cost function in this model is calculated from the following equation (13):

$$\text{Cost} = c1*(1-R) + c2* M \tag{13}$$

$$\text{Reliability } R(I) = \text{mean}(R(VMt1), R(VMt2)... R(VMtn)) \tag{14}$$

In the equation (13), M is makespan of countries and R is mean reliability of resources in countries obtained from equation (14). In equation (14), R(VMt1) is the reliability of vm where first task is to be executed, R(VMt2) is the reliability of vm where second task is to be executed, and so on. R(VMtn) is the reliability of vm where nth task is to be executed. c1 and c2 values are coefficients to prioritize the order of importance makespan or reliability factors and sum of these values must be 1. These values are determined by the user.

4) *Initial empires generation*: some countries with highest power are considered as empire and other countries as colonies. For this, countries are ordered as ascending on basis of their cost value. So, countries in the beginning of the array will have lower cost. Normalized cost and power of empires and number of their colonies is obtained from equations (1), (2), (3).

5) *Assimilation*: In this step, assimilation policy is done. In fact, distance of empire and colony is calculated from equation (15) and then colony new position is calculated from equation (16).

$$d = \text{Imperialist position} - \text{Colony position} \tag{15}$$

$$X = \text{round}(\text{Colony Position} + \alpha * \beta * \text{int}(\text{Rnd}(2)) * d) \tag{16}$$

In equation (16),  $\beta$  is a number greater than one and close to 2.  $\beta$  is made to colony country during move to empire country, close to it from different sides.  $\alpha$  is a desired parameter which increases increase search around empire and decreases to make colonies move close to Vector Empire to colony. In this model,  $\alpha$  is considered 0.5.

6) *Revolution*: Revolution means sudden changes in the position of a country that values change with a defined rate randomly. In fact, in this step, resources with random number are allocated to tasks that will a new country appear. Revolution rate in this model is 0.1. Figure 3 shows a sample of this change.

7) *Exchange colony and empire*: as explained before, if a colony gets lower cost while moving to empire then the colony and empire exchange positions.

8) *Empires Competitive and Eliminate Weak Empire*: Total power of empire is obtained from equation (4). To takeover colonies of other empires, first according equation (5), from total cost of empire, is defined as it's normalized total cost and then, probability of takeover each empire (depend on the power of the empire), with considering total cost of empire is calculated from equation (6).

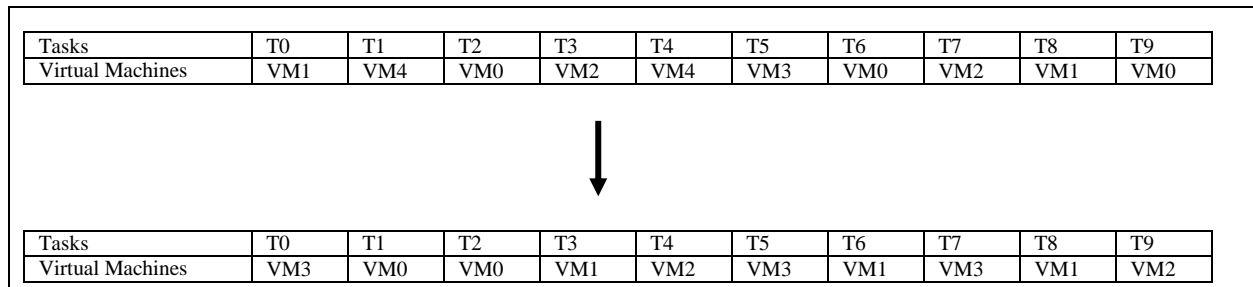


Fig. 3. Revolution

9) *No need to calculate CDF1*, make this mechanism to perform faster than roulette wheel in genetic algorithm. This step is ended when one of the empires takes over a colony, then after a few iteration, weakest empire will be without any colonies. In this case that empire will be removed from the empire list.

10) *Termination*: The algorithm continues until a converge

11) *Condition or reach to total number of iteration reached*.

The pseudo code of proposed ICA is given in Figure 4.

The flowchart of proposed method is given in Figure 5.

```

1 Initialize Tasks and virtual machines.
2 While (No. of countries < initial population size) {
3 Set the status of all tasks in set T to unscheduled to generate new country
4 While (Set T has unscheduled tasks) {
5 Select a task t from set T of tasks which has not been allocated to any virtual machine.
6 Select a virtual machine vm randomly from set VM of virtual machines.
7 Schedule task t on virtual machine vm.}
8 Store the new generated country in population P.}
9 Evaluate cost of all countries in population P by calling reliability evaluation algorithm.
10 Sort the countries in ascending order according to their cost value.
11 Generate initial empires from top list of countries and colonies as other countries.
12 While (total number of iteration < iteration threshold or No. of empire > 1)
{
13 Calculate distance of empire and colony and move colony.
14 If (Revolution rate > Revolution threshold) then
15 Apply Revolution.
16 If (empire cost > colony cost) then
17 Exchange Empire and colony position.
18 Calculate total power of empire.
19 Calculate normalized total power of empire.
20 Make vector D
21 Get max index of vector D
22 Give a colony of weakest empire to highest power empire.
23 If weakest empire (No. of colonies) <=1 then
24 Eliminate Weakest Empire and change to colony.}
25 Return Best Country with min Cost

```

Fig. 4. The pseudo code of proposed ICA

<sup>1</sup> Cumulative distribution function

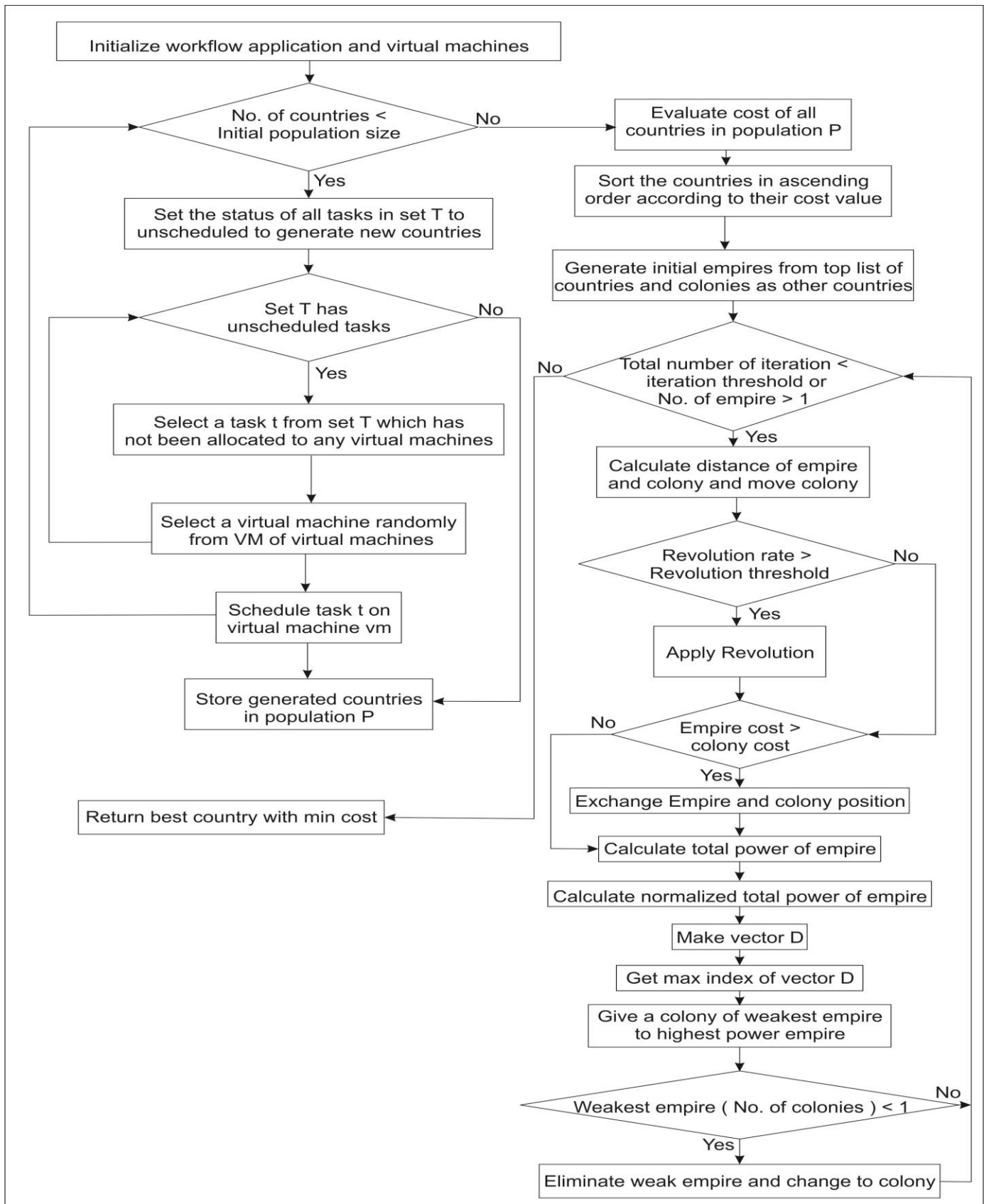


Fig. 5. The flowchart of proposed method

V. PERFORMANCE EVALUATION

To ensure proper functioning of the proposed method, results of the testing should be compared to other methods. Until now, different papers are provided in this field that often are handed makespan problem. In [21], the makespan and reliability are considered by using genetic algorithm. So, in this paper proposed algorithm is compared to genetic algorithm. CloudSim is used for implementation of the algorithm [24]. Table II gives information about simulation.

TABLE II. SIMULATION CONDITION

Parameters	Value
Number of Datacenters	2
Number of Hosts	2
MIPS Rate of Host	1000000
Ram of Host	6144
Number of VMs on each Host	2,3
MIPS Rate of VMs	2600,3200
Ram of VM	512
Mode of VMs on Host	Space-Shared
Total Number of VMs	5,10
Number of Tasks	50-250
Task size (No. of Instructions)	1000-2000

The experiment is performed on a computer with a dual-core, 2.4 GHz processor and 2GB of RAM. The test was done once on 5 virtual machines and another one on 10 virtual machines in an unreliable cloud environment. The scheduling algorithm estimates execution time of tasks by using MIPS rate of virtual machines. MIPS rates of some virtual machines are considered as 2600 and 3200. Instruction numbers of tasks are random numbers between 1000 and 2000. Table III and Table IV show the parameters in ICA and GA, respectively.

TABLE III. ICA PARAMETERS

Initial Population Generation	Randomly
Initial Population Size	100
Number of Iterations	100
Revolution Rate	0.1
Assimilation Coefficient	2
Assimilation Angle Coefficient	0.5
zeta ( $\xi$ )	0.02
RF Value	0.05

TABLE IV. GA PARAMETERS

Initial Population Generation	Randomly
Initial Population Size	100
Number of Iterations	100
Mutation operation	swapping
Mutation rate	0.01
Crossover operation	Single point
Crossover rate	0.5

Experiments are done for both algorithms with different number of tasks and resources. Table V shows average results of the experiment after running the algorithm 20 times for parameters such as makespan and reliability.

TABLE V. COMPARISON OF ALGORITHMS WITH DIFFERENT TASKS AND RESOURCES

Number of Tasks	Number of Recourses	GA		ICA	
		Makespan	Reliability	Makespan	Reliability
100	5	13.42	0.96	12.1	0.98
200	10	14.01	0.9	12.71	0.93

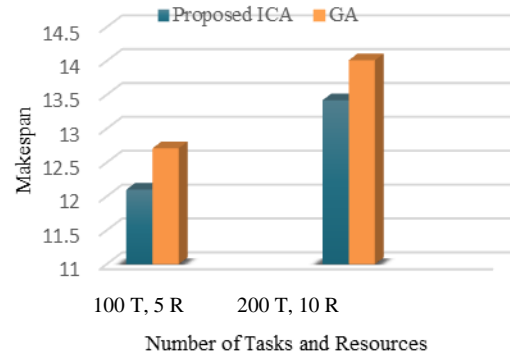


Fig. 6. comparison of makespan in both algorithm with increasing number of tasks and resources

The values of two results are compared together. The bar diagram in Figure 6 shows makespan on resources in both algorithms for 100 tasks with 5 resources and 200 tasks with 10 resources. As shown, the proposed method reduces makespan rather than genetic algorithm by 5 percent in the first experiment and 4.4 percent in the second experiment. The ICA algorithm has useful operations such as revolution and assimilation and help to better search the problem space but in genetic algorithm mutation is used to avoid local optimum. Hence, ICA can find optimum solution than GA

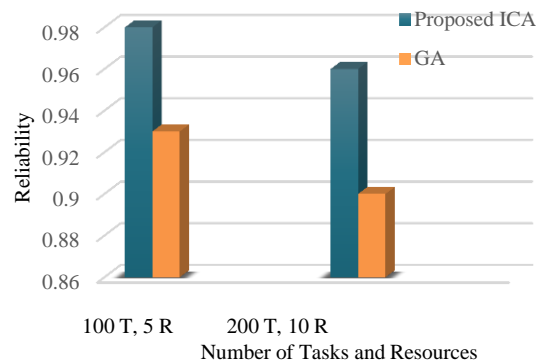


Fig. 7. compare reliability of both algorithm with increasing number of tasks and resources

Figure 7 shows reliability on resources in both algorithm for 100 tasks with 5 resources and 200 tasks with 10 resources. As shown, the proposed method increases reliability rather than genetic algorithm (in first experiment 5.4 percent and in second experiment 6.7 percent). Therefore,



according to this Figure and Figure 6, ICA proves to be able to better search the problem space rather than GA. In the other hands selected cost function here in linear condition has fitting to two objects separately.

For more studies, we use different environment scale. The results of makespan of both methods with different tasks and 5 resources are in accordance with Figure 8. In GA, during the initial stage of population, virtual machines with high reliability are selected. In this algorithm, makespan is less important. In addition, in imperialist competitive using vector D is used instead of roulette wheel that is quicker. So as shown in Figure 8, the proposed algorithm has lower makespan in comparison to GA.

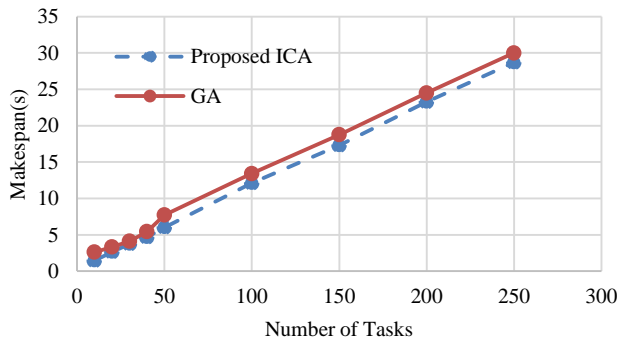


Fig. 8. comparison of makespan

Reliability evaluation results of VMs are shown in Figure 9. As shown, in low number of tasks, reliability of both algorithms is near to each other but by increasing number of tasks, reliability of proposed ICA will be higher than GA.

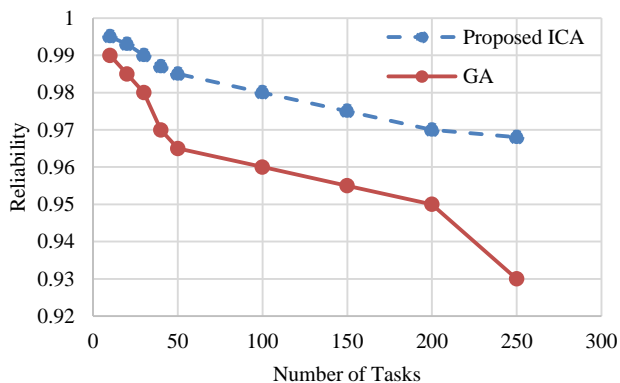


Fig. 9. comparison of reliability

## VI. CONCLUSION

In this paper, for resource allocation, makespan and reliability were used as a fitness function. In fact, two parameters were used in ICA as a cost function. In cross layer, initial reliability of all resources is equal to 1 and after they implemented tasks, the reliability increased or decreased according to successful or failure implementation. Then, by using imperialist Competitive algorithm different solutions are checked and their cost function is calculated by makespan and

reliability values. In comparison with other methods, results show that the proposed algorithm improves makespan and reliability rather than GA. In fact, the proposed method reduces makespan rather than genetic algorithm in experiment 100 tasks and 5 resources 5 percent and in in experiment 200 tasks and 10 resources 4.4 percent while is increase reliability in experiment 100 tasks and 5 resources 5.4 percent and in experiment 200 tasks and 10 resources 6.7 percent. Results show that ICA can search problem space better than GA.

## REFERENCES

- [1] Peter Mell, Timothy Grance: The NIST Definition of Cloud Computing, National Institute of Standards and Technology, Special Publication 800-145 (2011)
- [2] Zhao, L., Sakr, S., Liu, A., Bouguettaya, A.: Cloud Data Management, XIX, 202 p. 86 illus., 50 illus., Springer, 2014
- [3] Chunye Gong, Jie Liu, Qiang Zhang, Haitao Chen and Zhenggu Gong.: The Characteristics of Cloud Computing, 39th International Conference on Parallel Processing Workshops, 2010.
- [4] Yu, J., Buyya, R., Kotagiri, A.: Workflow Scheduling Algorithms for Grid Computing, vol. 146, pp. 173–214. Springer, Heidelberg (2008)
- [5] Wang, P.C., Korfhage, W.: Process Scheduling using Genetic Algorithm. In: Parallel and Distributed Proceeding Seventh IEEE Symposium, pp. 638–641 (1995).
- [6] Page, A.J., Naughton, T.J.: Dynamic Task Scheduling using Genetic Algorithm for Heterogeneous Distributed Computing. In: Proceedings 19th IEEE Conference on Parallel and Distributed Processing Symposium (2005).
- [7] Moattar, E.Z., Rahmani, A.M., Derakhshi, M.R.F.: Job Scheduling in Multiprocessor Architecture using Genetic Algorithm. In: 4th IEEE Conference on Innovations in Information Technology, pp. 248–251 (2007).
- [8] Mocanu, E.M., Florea, M., Ionut, M.: Cloud Computing Task Scheduling Based on Genetic Algorithm. In: System IEEE Conference, pp. 1–6 (2012).
- [9] Shekhar Singh, Mala Kalra, Task scheduling optimization of independent tasks in cloud computing using enhanced genetic algorithm-International journal of application or innovation in engineering & management (ijaiem)- volume 3, issue 7, july 2014.
- [10] Van den Bossche, R., Vanmechelen, K., Broeckhove, J.: —Cost Optimal Scheduling in Hybrid IaaS Clouds for Deadline Constrained Workloads. In: 3rd IEEE International Conference on Cloud Computing, Miami (July 2010).
- [11] Tasquia Mizan, Shah Murtaza Rashid Al Masud, Rohaya Latip.: Modified Bees Life Algorithm for Job Scheduling in Hybrid Cloud, International Journal of Engineering and Technology Volume 2 No. 6, June, 2012.
- [12] Yue Miao: Resource Scheduling Simulation Design of Firefly Algorithm Based on Chaos Optimization in Cloud Computing, International journal of Grid Distribution Computing, Vol.7, No.6 (2014), pp.221- 228.
- [13] Hany Seidgar, Mostafa Zandieh, Hamed Fazlollahtabar, Iraj Mahdavi.: Simulated imperialist competitive algorithm in two-stage assembly flow shop with machine breakdowns and preventive maintenance, Journal of Engineering Manufacture February 18, 2015.
- [14] Zhao, L., Ren, Y., Sakurai, K.: —A Resource Minimizing Scheduling Algorithm with Ensuring the Deadline and Reliability in Heterogeneous Systems!. In: International Conference on Advance Information Networking and Applications, AINA. (IEEE 2011).
- [15] Felix C. Gärtner, “Fundamentals of Fault-Tolerant Distributed Computing in Asynchronous Environments”, ACM Computing Surveys, Vol. 31, No. 1, March 1999.
- [16] Yilei Zhang, Zibin Zheng and Michael R. Lyu, “BFTCloud: A Byzantine Fault Tolerance Framework for Voluntary-Resource Cloud Computing”, 4th International Conference on Cloud Computing, IEEE, 2011.
- [17] Arvind Kumar, Rama Shankar Yadav, Ran vijay and Anjali Jain “Fault Tolerance in Real Time DistributedSystem”, International Journal on

- Computer Science and Engineering (ICSE), Vol. 3, ISSN: 0975-3397, No. 2, Feb 2011.
- [18] AnjuBala, InderveerChana, "Fault Tolerance- Challenges, Techniques and Implementation in Cloud Computing", International Journal of Computer Science (IJCSI) Issues, Vol. 9, Issue 1, No 1, January 2012.
- [19] Ravi Jhavar, Vincenzo Piuri and Marco Santambrogio, Member of IEEE, "Fault Tolerance Management in Cloud Computing: A System-Level Perspective", IEEE, 2012.
- [20] Anjali D.Meshram, A.S.Sambare, S.D.Zade: Fault Tolerance Model for Reliable Cloud Computing, International Journal on Recent and Innovation Trends in Computing and Communication, 600-603 (2013).
- [21] Lovejit Singh and Sarbjeet Singh: A Genetic Algorithm for Scheduling Workflow Applications in Unreliable Cloud Environment, Springer-Verlag Berlin Heidelberg, 139-150 (2014).
- [22] Rajkumar Buyya, Rodrigo N. Calheiros, Jungmin Son, Amir Vahid Dastjerdi, and Young Yoon : Software-Defined Cloud Computing: Architectural Elements and Open Challenges, Cloud Computing and Distributed Systems (CLOUDS) Laboratory Department of Computing and Information Systems, arXiv:1408.6891v2 [cs.DC] 19 Feb 2015.
- [23] E. Atashpaz-Gargari and C. Lucas: Imperialist Competitive Algorithm An algorithm for optimization inspired by imperialistic competition, IEEE Congress on Evolutionary Computation, 2007.
- [24] Calheiros, R.N., Ranjan, R., De Rose, C.A.F., Buyya, R.K.: CloudSim: A Novel Framework for Modelling and Simulation of Cloud Computing Infrastructures and Services. GRIDS Laboratory. The University of Melbourne, Australia (2009).

# Real-Time Gender Classification by Face

Eman Fares Al Mashagba  
Computer Sciences Department  
Zarqa University  
Zarqa, Jordan

**Abstract**—The identification of human beings based on their biometric body parts, such as face, fingerprint, gait, iris, and voice, plays an important role in electronic applications and has become a popular area of research in image processing. It is also one of the most successful applications of computer–human interaction and understanding. Out of all the abovementioned body parts, the face is one of most popular traits because of its unique features. In fact, individuals can process a face in a variety of ways to classify it by its identity, along with a number of other characteristics, such as gender, ethnicity, and age. Specifically, recognizing human gender is important because people respond differently according to gender. In this paper, we present a robust method that uses global geometry-based features to classify gender and identify age and human beings from video sequences. The features are extracted based on face detection using skin color segmentation and the computed geometric features of the face ellipse region. These geometric features are then used to form the face vector trajectories, which are inputted to a time delay neural network and are trained using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) function. Results show that using the suggested method with our own dataset under an unconstrained condition achieves a 100% classification rate in the training set for all application, as well as 91.2% for gender classification, 88% for age identification, and 83% for human identification in the testing set. In addition, the proposed method establishes the real-time system to be used in three applications with a simple computation for feature extraction.

**Keywords**—*Biometrics; Face Detection; Geometry-based; Gender Classification; Quasi-Newton Algorithms*

## I. INTRODUCTION

Automatic analysis of video data is a very challenging problem. To find a particular object in a video stream and automatically decide if it belongs to a particular class, one should utilize a number of different machine learning techniques and algorithms, as well as solve object detection, tracking, and recognition tasks [12].

Automatic human recognition tasks based on pattern recognition and artificial intelligence (AI) use different biometric body parts, such as face, fingerprint, gait, iris, and voice. Of all these body parts, the face is one of the most popular traits because of its unique features and impression [9]. Identity verification using face recognition can be performed without the cooperation and knowledge of the person being identified. However, recognizing a person becomes difficult because of the variations in pose, illumination, occlusion, expression, and so forth. [9]. Human identification using the face is critical because it considers multiple parts of the facial structure, and the face of a person changes with the passage of time. In fact, individuals can process a face in a variety of ways

to categorize it by its identity, along with a number of other demographic characteristics, such as gender, ethnicity, and age. In particular, recognizing human gender is important because people respond differently depending to on gender. In sum, a successful gender classification approach can boost the performance of many other applications, including person recognition and smart human–computer interfaces [20]. Face detection is critical to the final result in several applications, such as face processing (i.e., face, expression, gender classification, and gesture recognition), computer–human interaction, human crowd surveillance, biometrics, video surveillance, AI, and content-based image retrieval. It can be viewed as a preprocessing step for obtaining the object region [7] [12] [20] [21].

Recent research carried out face detection by using a color-based algorithm [5] [7] [10] [15] [18] [22] [24], where the segmentation of skin-colored regions becomes robust only when a proper color model is chosen. Several color models exist RGB, YCbCr, and HSV color models, and each has a specific work field and strength [7]. The RGB color space consists of the three additive primaries: red, green, and blue. The RGB model simplifies the design of computer graphics systems but is not ideal for all applications. The YCbCr color space was defined in response to increasing demands for digital algorithms in handling video information and has since become a widely used model in digital videos. Given that hue, saturation, and intensity value are properties used to describe color, a corresponding color model, HSV, logically exists. When using the HSV color space, knowing what percentage of blue or green is required to produce a color is unnecessary; the hue is adjusted to get the desired color [7].

The rest of this paper is organized: Previous Studies in section II, proposed methodology in section III, Experimental results in section IV, finally conclusion and future work in section IV.

## II. PREVIOUS STUDIES

The present study proposes a short review of the latest methods used in gender classification. In facial feature extraction, various methods apply to extract the features from image or video sequences: geometry-based, template-based, color-based segmentation, and appearance-based methods [6] [21] [22].

The geometry-based method extracts features using geometric information, such as the relative positions and sizes of the facial components. This method, however, requires the classifier to use a large number of features. The techniques proposed in [6] extract geometric features, such as left eye width, righteye width, nose width, left eye center to mouth left

corner, right eye center to mouth right corner, left eye center to mouth right corner, and right eye center to mouth left corner, mouth left corner to middle of chin, and mouth right corner to middle of chin. [17] proposed using geometric features, such as the distance between eyebrows to an eye, eyebrow to nose top, nose top to mouth, eye to mouth, left eye to right eye, nose width, and mouth width, all of which are extracted using the Viola–Jones algorithm. [1] [8] [15] applied a combination of global and grid features. Global features include inter-ocular distance, the distance between the lips to the nose tip, the distance between the nose tip to the line joining the two eyes, the distance between lips to the line joining the two eyes, eccentricity of the face, ratio of dimension, and width of the lips. Grid features include skin color, moustache region, lip region, eye tail, forehead, eyelid, and nose wing of the face image.

The template-based approach matches facial components to previously designed templates using an appropriate energy functional. The best match of a template in the facial image will yield the minimum energy; this approach can be too complex due to the extensive computation involved and is only effective when query and model images have the same scale, orientation, and illumination properties. Color-segmentation makes use of skin color to isolate the face and any non-skin color region. In this approach, the quality of the image coupled with illumination and hue plays an important role in image recognition and rate of classification. For a complicated background, [18] based their face detection method on the skin color feature.

The features in the appearance-based method differs from the simple facial features, such as eyes and mouth, used in other approaches. Any extracted characteristic of the image refers to a feature. In methods such as principal component analysis (PCA), which was proposed by [5] [14] [19] [21] [24] [27], a face image is represented as a two-dimensional  $N$  by  $N$  array of intensity values or a vector of dimension  $N^2$ . Then PCA finds an  $M$ -dimensional subspace whose basis vectors correspond to the maximum variance direction in the original image space. New basis vectors define a subspace of face images called face space. All images of known faces are projected onto the face space to find sets of weights that describe the contribute of each vector. The face can be identified through a comparison of a set of weights for the unknown face and the sets of weights of known faces. [12] applied an adaptive feature generation algorithm trained by means of the optimization procedure according to the LDA principle. Gabor wavelets were also used to extract the feature vector. These approaches are commonly used for facial recognition rather than person identification. [3] [24] applied Gabor filters at five scales and eight orientations for detecting skin regions. In [13], a Fourier–Gabor filter was applied to extract features from the face images. A different size, orientation, and scale values were adopted.

Regardless of the above mentioned methods, a variety of classification techniques can be used for recognition, such as decision trees, neural networks, nearest neighbors classifier (NNC), support vector machines (SVMs), Bayesian networks, fuzzy logic, and genetic algorithms, among many others. [11] trained and tested three classifiers SVM, back propagation NNs

(BPNNs), and KNN. These classifiers are then optimized through the GA; using this approach; they got promising results in terms of the classification error rate and the minimization of computation time. [6] proposed a variant of the decision tree algorithm for gender classification of frontal images owing to its distinctive features. Their technique showed robustness and relative scale invariance for gender classification. [22] got high accuracy by using SVMs for gender classification. [3] [21] [26] produced very promising recognition rates for three applications face recognition, facial expression recognition, and gender classification, and reasonable results in all databases with the same set of features and (NNC) classifiers. The system also had real-time capability and was automatic. [23] used an unsupervised learning technique to classify DCT-based feature vectors into groups and identify if the subject in the input image was “present” or “not present” in the image database. After training for approximately 850 epochs, the system achieved a recognition rate of 81.36% for 10 consecutive trials. The main advantage of this technique is its high-speed processing capability and low computational requirements in terms of speed and memory utilization. In [14], PCA matrixes with different numbers of components (40, 50, and 60) were passed to the NN, which conducted the framework with a different layer number, different learning algorithms, and a different number of neurons in each layer, and achieved a high correct classification rate of approximately 83.5%. [8] used posteriori class probability and Artificial Neural Network to classify gender and age, respectively. The results achieved were 100% for face recognition, 98% for gender classification, and 94% for age classification. In [18], the recognition step was applied by SVM. A good performance of the gender classification test was also achieved on a relatively large-scale and low-resolution video database. [17] applied the Artificial Neural Network for face, facial expression, and gender classification, and produced reasonable results in all databases. [12] proposed using SVM and presented the experimental results gained on a large image dataset. More than 90% accuracy of the viewer's gender recognition was achieved.

The myth of a successful real-time gender classification depends greatly on the right choice of features and classification method. In this paper, an algorithm is introduced to enhance real-time gender classification. The face region is detected using skin color segmentation with HSV color space, and the features are extracted using global geometric properties. A Time Delay Neural Network (TDNN) algorithm is used for classification tasks. The suggested method is implemented by using our own datasets.

### III. PROPOSED METHODOLOGY

The steps of the proposed method can be summarized as follows:

- 1) *Extract video frames.*
- 2) *Delete abnormal images from the start and the end manually.*
- 3) *Detect the face region from each frame using skin color segmentation.*

- 4) Extract features from each image using the geometry-based approach.
- 5) Classify gender using dynamic neural networks.

#### A. Face detection

With our method, a dependable result can be achieved for video face segmentation based on skin color. The quality of face detection is critical to the final result of the whole system as an imprecise determination of face position can lead to wrong decisions at the recognition step [12].

Each video sample is taken from our own database. The segmentation of skin-colored regions becomes robust only by choosing proper color model. Numerous applications use the HSV color model. Machine vision uses HSV color space to identify the color of different objects. Image processing applications, such as histogram operations, intensity transformations, and convolutions operate only on an intensity image. These operations are performed with much ease on an image in the HSV color space [2] [7] [12]. We applied skin color segmentation with the HSV color model to separate all skin color pixels in connecting components that use an adaptive Gaussian mixture model (GMM) proposed by [10]. The adaptive GMM can adapt the model parameters to cope with changing imaging conditions, such as lighting and noise. The segmentation method detected the ellipse that represents the skin region, both original and skin image, as shown respectively in Figure. 1(a) and (b).



Fig. 1. (a) Initial Image (b) Extracted Region

#### B. Feature Extraction

Geometric-based features are extracted using ellipse mathematical definition and properties [28]. Global features are calculated for a video sequence frame from the ellipse for the face region, horizontal center ( $X_0$ ), and vertical center ( $Y_0$ ) of the detected face, where  $a$  and  $b$  are the semi-major and semi-minor axes (half of the major and minor axes of the ellipse) respectively. The area enclosed by an ellipse area is shown as Equation (4). The two foci (the term “focal points” are also used) of an ellipse are two special points  $F_1$  and  $F_2$  on the ellipse's major axis and are equidistant from the center point. The sum of the distances from any point  $P$  on the ellipse to those two foci is constant and equal to the major axis ( $PF_1 + PF_2 = 2a$ ), denoted as eccentricity as in Equation (3). The eccentricity of an ellipse, usually denoted by  $\epsilon$  or  $e$ , is the ratio of the distance between the two foci to the length of the major axis or  $e = 2f/2a = f/a$ . The ratio of dimension and four-quadrant inverse tangent (arctangent) of the real parts of the semi-major axis ( $a$ ) and semi-minor axis ( $b$ ) are calculated using Equation (2). Figure 2 shows the geometry of the face region. The distance from the center  $C$  to either focus is  $f = ae$ ,

which can be expressed in terms of the major and minor radii as shown in Equation (1):

$$f = \sqrt{a^2 - b^2} \tag{1}$$

$$P = \text{atan2}(a, b) \tag{2}$$

The eccentricity of the ellipse (commonly denoted as either  $e$  or  $\epsilon$ ) is

$$e = \epsilon = \sqrt{\frac{a^2 - b^2}{a^2}} = f/a \tag{3}$$

The area enclosed by an ellipse is

$$\text{area} = \pi ab \tag{4}$$

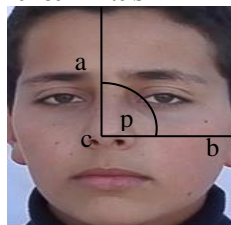


Fig. 2. Geometry of the Face Region

The features used to track the difference between each face region for given adjacent frames in the sequences and the resulting spatio-temporal vectors input in the classification step.

#### C. Classification Method

Finally, a TDNN is used for gender classification. The TDNN is a well-known classifier that is very successful in classifying spatio-temporal patterns as described in details in my previous research [16] and is thus used in this research. The architectural characteristics of the TDNN have an additional feature, known as a tapped delay line in its input neuron [12]. In the present research, the delays are: in the input layer,  $d_1 = 0:3$ ; in two hidden layers,  $d_2 = 0:3$  and  $d_3 = 0:5$ ; and in the output layer,  $d_4 = 0:5$ . Training and testing require a sufficient number of video databases. There are many commonly used databases for the tasks of human face recognition such as Multi-PIE database, the FERET database, and SCface, but they donot contain a sufficient number of videos of different individuals and are inadequate for our applications. We collected our own database. Table1 describes the frame parameters.

TABLE I. PARAMETERS OF THE PROPOSED TRAINING AND TESTING IMAGE DATABASE

Parameter	Value
Number of males	5
Minimum image resolution	243×360
Total number of images	2255
Number of females	5
Color space format	RGB
People's age	8–60 years old
Lighting conditions, background, and facial	No restrictions

expression	
------------	--

This dataset was split into two independent video sets: training and testing. The training set was used for feature generation and TDNN classifier construction. A performance evaluation of the trained classifier was approved with the use of the testing set.

Various batch training algorithms can train a network. Three types of training algorithms have eight training functions: gradient descent algorithms (traingd, traingdm, trainrp), conjugate gradient algorithms (traincg, traincgf, traincgp), and quasi-Newton algorithms (trainbfg, trainlm) according to the explanation in [25]. dependent on the performance results, we used quasi-Newton algorithms (trainbfg). Determining an appropriate NN architecture depends greatly on the experience of the person controlling the experiment [16]. The feature vector for gender recognition by the network has to be encoded and placed in the input nodes of the network. These input values must be scaled so they are in the same range which can help save time and get high performance of the training methods [16].

#### IV. EXPERIMENTAL RESULTS

To evaluate the performance of the gender classification algorithm, we prepared our own video sample for the subject. Figure 3 shows sample face sequences from the dataset. TDNN with conjugate gradient algorithms, BFGS quasi-Newton backpropagation (trainbfg) training function in our experiments, The figure also shows the performance result of the human identification and gender and age classification system.



Fig. 3. Face Sequences from the Dataset

All videos in the experiments have a range of 100–240

Application Type	Training Set Classification Rate	Testing Set classification Rate
Gender classification		91.2
Age classification	100%	88.0
Human identification		83.0

frames. We collected 10 video samples from each person and used MATLAB 10 to implement the development process of gender classification. For each frame, global geometric features from the face segment part were extracted to form the face vector trajectories. These trajectories include vertical center, horizontal center, semi-major axis, semi-minor axis, and four-

quadrant inverse tangent (arctangent) of the real parts of Y and X for each image sequence.

The frames that did not display a face part from the samples were eliminated. The feature vector of all frames form an input vector that trains the TDNN, which takes 70% of the samples for training and 30% for testing of all persons. During this research, the TDNN architecture was applied as shown in Figure. 4. Based on previous research [16], the sigma mode function “tansig” is used in the input and hidden layers, whereas the linear transfer function pure linear “purelin” is used in the output layer. We created 4 layers of TDNN with 5 input neurons and 20 hidden neurons, with the delay of 0–3 for the first two layers and 0–5 for another layer.

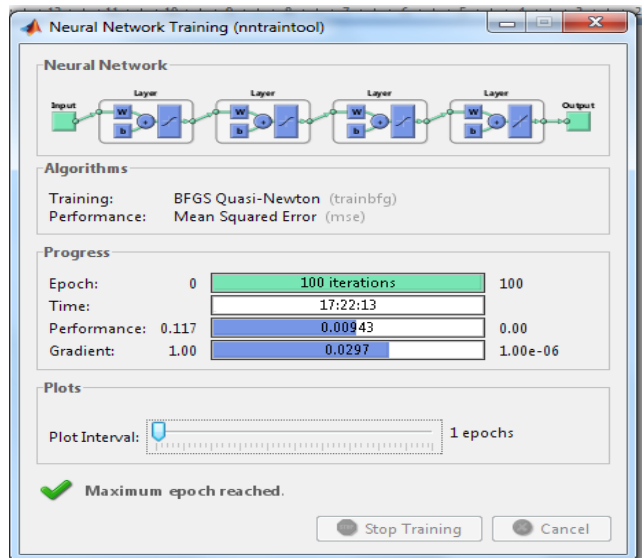


Fig. 4. TDNN Architecture

The performance results with 200 epochs of gender classification, human identification and age classification are shown in Figure.5a, b, and c, respectively.

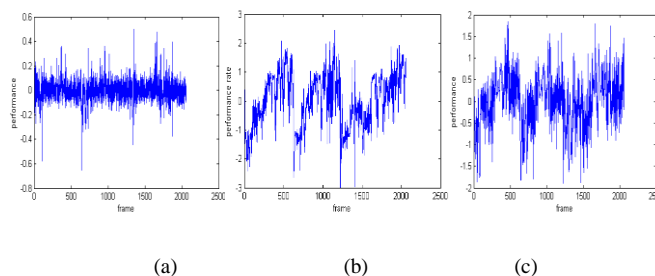


Fig. 5. Mean Squared Error of (a) Gender Classification, (b) Human Identification, and (c) Age Classification

TABLE II. CLASSIFICATION RATE USING 70% FOR TRAINING AND 30% FOR TESTING

Table 2 lists the experiments results based on classification rate and application purposes. The conjugate gradient algorithms (trainbfg) achieved promising results in training the TDNN [16]. The results were sorted based on the time consumed and the classification rate, where the highest classification rates in testing sets are in gender classification,



age classification, and human identification. This result implies that gender, age, and human beings can be classified with promising recognition rate by using global geometric features from video. The developed method based on global geometric facial features achieves a high classification rate of 100% in the training set for all application, as well as 91.2% for gender classification, 88% for age classification, and 83% for human identification in the testing set. The proposed method also establishes the real-time system to be used in three applications with a simple computation for feature extraction. During the experiment, we attempted to train 30 frames from each sample and achieved the same result from all frame trainings. The results show that the suggested methods with all frames and with a sample sequence of frames outperform state-of-the-art methods.

## V. CONCLUSION AND FUTURE WORK

A real-time classification system using TDNN is described according to BFGS training algorithms and used for gender and age classification and human identification. The developed system implies that we can classify gender, age, and human being with promising recognition rate by using global geometric features from video. The developed method based on global geometric facial features achieves a high classification rate of 100% in the training set for all application, as well as 91.2% for gender classification, 88% for age classification, and 83% for human identification in the testing set. Future work will be devoted to implementing and examining the other features extracted from face parts by various techniques with real-time application.

## ACKNOWLEDGMENT

This research was funded by the Deanship of Scientific Research in Zarqa University, Jordan.

## REFERENCES

- [1] Alok Verma, S Achyut Rajy, Abhishek Midyaz, Jayasree Chakraborty,(2014). Face Detection using Skin Color Modeling and Geometric Feature. Conference: 2014 International Conference on Informatics, Electronics & Vision (ICI) ). Dhaka: IEEE, 2, 14.Pp. 1-6. <http://dx.doi.org/10.1109/ICIEV.2014.6850755>
- [2] Erno Mäkinen, Roope Raisamo. (2008). an Experimental Comparison of Gender Classification Methods. Pattern Recognition Letters, 1544–1556. <http://dx.doi.org/10.1016/j.patrec.2008.03.016>
- [3] Md. Hafizur Rahman, Suman Chowdhury, Md. Abul Bashar. (2013). an Automatic Face Detection and Gender Classification from Color Images using Support Vector Machine. Journal of Emerging Trends in Computing and Information Sciences, 5-11.
- [4] Meltem Demirkus, Matthew Toews, James J. Clark, Tal Arbel. (2010). Gender Classification from Unconstrained Video Sequences. IEEE, 55-62. <http://dx.doi.org/10.1109/cvprw.2010.5543829>
- [5] Ms. Rutuja G. Shelke, Prof. S. A. Annadate. (2014). Face Recognition and Gender Classification using Features of Lips. International Journal of Innovation and Scientific Research, 337-341.
- [6] Muhammad Naem Ahmed Khan, Sheraz Ahmed Qureshi and Naveed Riaz. (2013). Gender Classification With Decision Tree. International Journal of Signal Processing, Image Processing and Pattern Recognition, 165-176.
- [7] Prashanth Kumar G. And Shashidhara M. (2014). Real Time Detection and Tracking of Human Face using Skin Color Segmentation and Region Properties. International Journal of Signal Processing Systems, 102-107.
- [8] Ramesha K, K B Raja, Venugopal K R And L M Patnaik. (2010). Feature Extraction Based Face Recognition, Gender and Age Classification. International Journal on Computer Science and Engineering, 14-23.
- [9] Reecha Sharma, M.S. Patterh. (2015). Indian Face Age Database: a Database for Face Recognition with Age Variation. International Journal of Computer Applications (0975 – 8887), 21-27.
- [10] Reza Hassanpour, Asadollah Shahbahrani, and Stephan Wong. (2008). Adaptive Gaussian Mixture Model for Skin Color Segmentation. Proceedings of World Academy of Science, Engineering and Technology, 1-6.
- [11] Sajid Ali Khan, Muhammad Nazir, Naveed RIAZ. (2013). Optimized Features Selection for Gender Classification using Optimization Algorithms. Turkish Journal of Electrical Engineering and Computer Sciences, 1479 - 1494. <http://dx.doi.org/10.3906/elk-1203-51>
- [12] Vladimir Khryashchev, Lev Shmaglit, Andrey Shemyakov, Anton Lebedev, (2014). Gender Classification for Real-Time Audience Analysis System. Open Innovations Association FRUCT, Proceedings Of 15th Conference of Fruct Association, St. Petersburg: IEEE, 2014, Pp. 52 - 59. <http://dx.doi.org/10.1109/fruct.2014.6872428>
- [13] Daoud, E. A. (2015). Face Recognition Fuzzy Clustering and Kernel Least Square. Journal of Computer and Communications, 2015, 1-7. <http://dx.doi.org/10.4236/jcc.2015.33001>
- [14] Ahmad Awwad, Ashraf Ahmad, Walid A. Salameh. (2013). Arabic Race Classification of Face Images. International Journal of Computers and Technology, 234-239.
- [15] Akshay Kajekar, Shreyash Patel, Sweta Chauhan, Prashasti Kanikar. (2012). Blending of Face Segments. International Journal of Computer Science and Information Technologies, 3929-3933.
- [16] AL Mashagba, E. F. (2015). Human Identification by Gait Using Time Delay Neural Networks. Computer and Information Science, 56-63. <http://dx.doi.org/10.5539/cis.v8n4p56>
- [17] Anushri Jaswante. Asif Ullah Khan. Bhupesh Gour. (2013). Gender Classification Technique Based on Facial Features using Neural Network. International Journal of Computer Science and Information Technologies, 839-843.
- [18] Chuanxu, W. (2011). Face Segmentation Based on Skin Color in Complicated Background and its Sex Recognition. Journal of Software, 1209-1216.
- [19] Dewi Agushinta R., Adang Suhendra, Sarifuddin Madenda, Suryadi H.S. (2010). Face Component Extraction using Segmentation Method on Face Recognition System. Journal of Emerging Trend in Computing and Information Sciences, 67-72.
- [20] Federico Matta, Usman Saeed, Caroline Mallauran and Jean-Luc Dugelay, (2008) Facial Gender Recognition using Multiple Sources of Visual Information. Multimedia Signal Processing, 2008 IEEE 10th Workshop On, Cairns, Qld: IEEE, 2008, Pp. 785-790. <http://dx.doi.org/10.1109/MMSP.2008.4665181>
- [21] Hlaing Htake Khaung Tin. (2012). Perceived Gender Classification From Face Images. I.J. Modern Education and Computer Science, 8-12.
- [22] Islaml, M. S. (2013). Gender Classification using Gradient Direction Pattern. Sci. Int (Lahore), 797-799.
- [23] Jawad Nagi, Syed Khaleel Ahmed, Farrukh Nagi. (2008). A MATLAB Based Face Recognition System using Image Processing and Neural Networks. 4th International Colloquium on Signal Processing and its Applications, 83-88.
- [24] Md. Hafizur Rahman, Tonmoy Das, Manamatha Sarnaker. (August 2013). Face Detection and Sex Identification from Color Images using Adaboost with SVM Based Component Classifier. International Journal of Computer Applications, 0975 – 8887.
- [25] Sanjeev Sharma, Ritu Tiwari, Anupam Shukla And Vikas Singh. (2011). Identification of People using Gait Biometrics. International Journal of Machine Learning and Computing, Vol. 1 (No. 4), 409-415.
- [26] Swaminathan, A. (2000). Gender Classification using Support Vector Machines., 2000. ENEE633: Statistical and Neural Networks pattern Recognition, 2000, (Pp. 306 - 311). <http://dx.doi.org/10.7763/IJMLC.2011.V1.61>
- [27] Xiaodong Li, Wei Jiang, Weijie Mao, Jian Li And Kai Chen. (2013). Gender Classification Based on Evolutionary Extreme. International Journal of Innovative Computing, Information and Control, 3839-3849.
- [28] Young, Cynthia Y. (2010). "Chapter 9". Precalculus. John Wiley and Sons. p. 831. ISBN 0-471-75684-9

# An Accelerated Architecture Based on GPU and Multi-Processor Design for Fingerprint Recognition

Mossaad Ben Ayed

College of Science and Humanities at AlGhat, Al Majmaah  
University  
Kingdom of Saudi Arabia

Sabeur Elkosantini

College of Engineering, King Saud University  
Kingdom of Saudi Arabia

**Abstract**—Fingerprint recognition is widely used in security systems to recognize humans. In both industry and scientific literature, many fingerprint identification systems were developed using different techniques and approaches. Although the number of conducted research works in this field, developed systems suffer for some limitations partially those related the real time computation and fingerprint recognition. Accordingly, this paper proposes a reliable algorithm for fingerprint recognition based on the extraction and matching of Minutiae. In this paper, we present also an accelerated architecture based on GPU and multi-processor design in which the suggested fingerprint recognition algorithm is implemented.

**Keywords**—Minutia; Fingerprint; Architecture design; recognition; Gabor filter; MPSOC

## I. INTRODUCTION

Individual identification presents a challenge for the modern society. In this context, biometric recognition presents the most popular method for identification. Particularly, fingerprint technique is the most used in industries for many reasons: cheap, secure and easy to deploy. Fingerprints are rich in details and contain a different form based on ridges. These forms define many characteristics point named "Minutia". Each individual has unique repartition of Minutia which is different than others. Consequently, fingerprint is always used in many systems as the identifier of humans. Minutia is defined as a local ridge characteristic. Fingerprint contains various types of Minutia, but usually two types of Minutiae are used: Termination and Bifurcation. Termination is defined as the end point of a ridge. Bifurcation is defined as the point where a ridge merges or splits into branch ridges [1].

### A. Fingerprint recognition process

In the scientific literature, the process of fingerprint recognition is always divided to different phases including: the pre-processing, the extraction of the Minutiae and the matching (see Fig. 1).

The objective of the first phase, the pre-processing, is applied on gray-scale images essentially to improve and divide fingerprints ridges from the background texture. The first step is to apply a filter algorithm. This step is important to ameliorate the quality of the image and to decrease noise [2]. Then the binarization step transforms the image into binary and lets the separation between ridges and background easier. The last step in the pre-processing step is the skeletonization which is based on thinning algorithms. The aim of this step is to thin ridges to only 1 pixel wide. This preserves the essential

information (Minutiae) with low size of storage. Also thinning algorithms reduces the data that represents Minutiae and make the treatment more effective and faster. In the scientific literature, there are many iterative methods for thinning including sequential [49] or parallel [3]. In this step, the most used window size is a 3X3 pixel. In this case the central black pixel has 8 neighbours that can be considered [3].

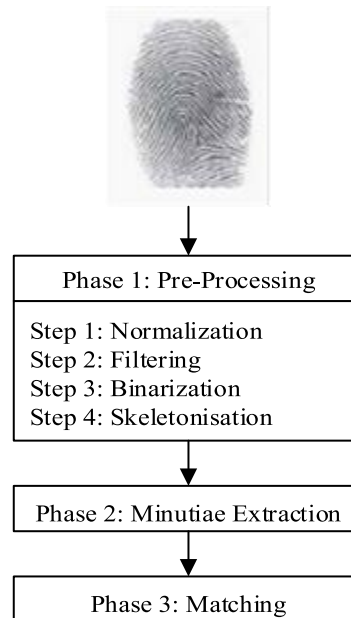


Fig. 1. Fingerprint recognition phases

The second phase is the extraction of the Minutiae. Many extraction methods were developed including those based on the nearest neighbour pixels around the central pixel [4]. Another method of extraction Minutia is presented in [5] which search the characteristics point Minutiae using thinned ridges. Other algorithms were based on classifier techniques [20]. [6, 7] tried to detect Minutiae using the ridge line without apply the thinning algorithms. In these works, different rules and ad-hoc methods are used to handle problems met on extraction. The extraction method proposed in [8] is based on Data-driven Error Correcting Output Coding (DECOC) classifier. This method presents a many advantages with other methods used. However, the outputs results depend not only on selected extraction algorithm but also on the pre-processing phase essentially binarization and skeletonization.

Matching phase aims to carry out fingerprint verification

and their generalization to find the factor of similarity between the input image and the corresponding database. This phase still suffers from the increasing of the time consumption, especially when the number of saved fingerprint is huge [42]. Although the number of the conducted research in the field of Minutiae matching for acceleration [48], many proposed architectures are not suitable for large database to maintain precision and the real time decision.

### B. Problem statement

As mentioned bellow, the pre-processing and the extraction Minutiae are an important phases to improve not only the quality of the image but also the recognition result. Accordingly, many research works were conducted to develop techniques and algorithms. Unfortunately, in spite of the number of research works in this field, there is still a lack for systems allowing the real-time fingerprint recognition. In [50] [51], the authors propose matching algorithm based on fingerprint database. But these algorithms let essentially sharing data with database and not suitable for real time identification for large database.

In this context, the objective of this work is the development of a new accelerated architecture based on GPU and multi-processor design based on a new algorithm for fingerprint recognition based on the extraction and matching of Minutiae.

The remainder of this paper is organized as follows: Section 2 introduces the developed pre-processing algorithm. Section 3 presents the proposed Minutiae extraction based on DECOC classifier. Section 4 describes the Minutiae matching phase and presents the adopted method. Section 5 presents the implementation of the suggested system and the obtained results. The paper is concluded by a conclusion and suggests some future works.

## II. PRE-PROCESSING PHASE

During fingerprint recognition, the image pre-processing represents an essential phase. The collected images from fingerprint sensor are challenged by the quality of fingerprinted person due to many conditions as skin condition, collection conditions, and the environment.

This phase can be divided into four steps: Normalisation, Filtering, binarization and skeletonization.

### A. Normalization step:

Normalization step can be defined as a method to reduce the diversification degree of the grayscale image captured by the sensor. As known, the fingerprinted is composed by ridges and valleys that form the structure of texture information. The aim of the normalization step is to facilitate the pattern capture and the fingerprint frequency. But this step still suffers from confusing between ridges and valleys [24].

The normalized step is based on equation (1).

$$G(i, j) = \begin{cases} M_0 + \kappa \times VAR \times |I_{(i,j)} - M| & I_{(i,j)} > M \\ M_0 - \kappa \times VAR \times |I_{(i,j)} - M| & Others \end{cases}$$

(1)

### B. Filtering step: Gabor

Filtering step is based on Gabor filter [13], [14], [21]. This filter can represent the local frequencies as mentioned in equation (2). It has both orientations: spatial domain and frequency domain. Developer can separate fingerprints from the background using Gabor filter algorithm based only on spatial domain. Indeed, the related works use only spatial domain because the Gabor filter take much time.

In this paper, spatial and frequency domain is used ensure the best filtering results [9]. This choice is adopted due to the proposed architecture that is based on GPU accelerator.

Gabor filter has the following equation:

$$h(x, y, \varphi, f) = e^{-\frac{1}{2} \left( \frac{x_\varphi^2}{\delta_x^2} + \frac{y_\varphi^2}{\delta_y^2} \right)} \cdot \cos(2\pi f x_\varphi)$$
$$avec \begin{cases} x_\varphi = x \cos \varphi + y \sin \varphi \\ y_\varphi = -x \sin \varphi + y \cos \varphi \end{cases} \quad (2)$$

where

f : frequency of the sinusoidal of the plane wave

$\varphi$  : orientation

$\delta_x, \delta_y$ : space constants of the Gaussian envelope along x and y axes, respectively.

### C. Binarization

Binarization step convert image from grayscale image to binary image. The most used method of binarization [15] is based on global threshold that consists in calculating a unique threshold for the total image.

The disadvantage of this method is shown when the fingerprint has different quality. In this case, binarization algorithm based on global threshold will eliminate many parts of the finger image (Figure 2.a).

This paper proposes a binarization method based on local threshold (Figure 2.b). This method is described as follows:

1) Divide the fingerprinted image into masks of 10 x 10 pixels.

2) Calculate the threshold of every mask. The last is determined using the equation (3).

$$Threshold = \frac{\sum M(i, j)}{nl * nc}$$

Where B(i, j) is the image

L \* C is the number of pixels (3)

This method shows good results especially when fingerprints and background will be separated. This method is more adopted in the field of fingerprints because the partition of the pixel intensity is not homogeneous.



Fig. 2. Binarization methods

#### D. Skeletonization

Skeletonization method is defined as a thinning algorithm that can be transforming the thickness of ridges into a single pixel [24].

The method used for skeletonization [16] is based on scanning the image by 3 x 3 blocks neighbourhood [17]. When this method is applied for the total of the image, the process costs too much time.

At first, this paper proposes a modified thinning algorithm like [17] to reduce the execution time of skeletonization step. For this reason, we consider that the process of thinning is

applied only when the 3 x 3 bloc contains more than two black pixels. An experimental result shows that the call of thinning function is reduced on third. Figure 3.a shows the results of skeletonization process.

As observed by figure 3.a, the zigzag ridge presents the major problem of skeletonization. It can influence the detection of the Minutiae because the changing of one pixel can modify the kind of Minutiae.

As a second contribution in skeletonization, smoothing filter is applied to decrease the zigzag effect, see figure 3.b.

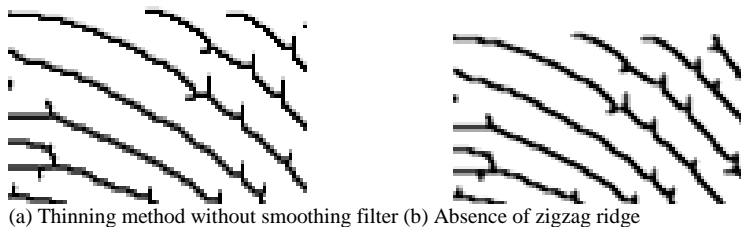


Fig. 3. Skeletonization method without (a) / with (b) Smoothing filter

### III. MINUTIAE EXTRACTION PHASE

Minutia extraction presents the second phase. This phase aims to find a particular point or feature in fingerprinted image named as Minutia. The last can be defined as the intersection of ridges or the ending of ridges. Minutia can be classified as two parts: (1) Bifurcation Minutia (intersection of two ridges) (2) Termination Minutia (ending ridges).

#### A. Discussion

Minutia extraction is based essentially on classification stage that aims to detect features [25]. There are many related work that propose algorithm to identify the Minutiae grayscale. In [26], Minutia is determined using a mask sized 3 x 3 bloc. A Minutia is identified if the pixel has one or more than two neighbors. But when the quality of fingerprinted image is poor is extremely unreliable. There many work that lets Minutia extraction by using different classifier as SVM [36], Bayes [37], Neuron, Fuzzy, etc. Table 1 resumes the most important contributions.

TABLE I. COMPARISON BETWEEN DIFFERENT METHODS FOR MINUTIAE EXTRACTION

	FAR	FRR
HAO GUO method [32]	4.18%	9.93%
OMER SAEED method [33]	1,12%	Not indicated
Ying HAO method [34]	1%	2.5%
Jiong Zang method [35]	0.04%	1.31%
Ching-Tang and al. method [36]	0.5%	0%
<b>Mossaad and al. method [27]</b>	<b>0%</b>	<b>0.02%</b>

In this paper, the DECOC classifier is used for Minutia extraction based on comparison in [27]. This classifier is adopted in many fields such as manuscript recognition. For this reason it is applied for fingerprint recognition. In the next section, a brief definition for the adopted method is presented.

#### B. Error Correcting Output Coding (ECOC)

In this section, Error correcting output code algorithm will be described. The DECOC classifier is based on ECOC classifier. The last have been used in different fields as network communication and information theory. The main purpose of the use is the reliability of transmitting binary signals and the integrity information. The idea is to add the superfluous parity bits to an information word. The novel word result is named code word. The code word presents a binary code string. Then, the ECOC algorithm calculates the distances between two code words using Hamming distance, as shown in equation (4). The last is determined by the count of the different bits in the two patterns.

*Hamming distance:*

$$y = \arg_{\min(j)} H(w_j, w(x)), j = 1, \dots, J \quad (4)$$

where  $w_j$  is the ideal code word assigned to group  $j$ .

$H(w_j, w(x))$  presents the main function that computes the Hamming distance between  $w_j$  and  $w(x)$ .

Data-driven presents an extending version of ECOC. The aim is to provide new solutions to the problem of multi-class. In this fact, DECOC compared with ECOC will be applied to two multi-class pattern recognition problems [8].

#### C. Methodology: Data-driven ECOC

Data-driven ECOC (DECOC) chooses the code words utilizing specific information from the training data. For a K-class problem as pair wise coupling, a decomposed algorithm is proposed. Also a  $K*(K-1)/2$  base learners are always needed.

This measure determines the learner that will be included in the ensemble. This measure is called confidence score. DECOC classifier is based essentially on two parameters: Separability criteria and confidence score [8].

- *Separability Criterion*

$$S(G) = \begin{cases} \frac{2}{|G|^2 - |G|} \sum_{0 \leq j < k < G} d(c_j, c_k) & |G| \neq 1 \text{ and } |G| \neq K-1 \end{cases} \quad (5)$$

- *Confidence Score*

$$C(f) = \begin{cases} \frac{S(G_{+/-}(f))}{S(G_+(f)) + S(G_-(f))}; & |G_+| \neq 1 \text{ and } |G_+| \neq K-1 \end{cases} \quad (6)$$

As mentioned before, the use of DECOC classifier is to make decision about the kind of pixels blocks. This classifier can be defined into two parts as follow:

```

/***** Calculate inter-class distance *****/
int *Inter_class_distance(int TD[x][y])
{
    for (i=vli;i<vli+5;i++)
        for (j=vci;i<vci+5;i++)

            *Dist=Hamming(TD[i][j],TD[i+5][j]);
    return Dist;
}
/***** Calculate separability criteria *****/
int *Separability_criteria(int *Dist)
{
    for (i=1;i<len;i++)
        sum[i]+=Dist[i];
    for (i=1;i<len;i++)
        if (condition 1)
            SC[G1][i]=Separability_Factor * sum[i];
        if (condition 2)
            SC[G2][i]=Separability_Factor * sum[i];
        if (condition 3)
            SC[G3][i]=Separability_Factor * sum[i];
    return SC;
}

```

Code 1: Inter-class distance and separability criteria functions

- Part 1: Find the best base learners based on higher confidence score. At first, the interclass distance and separability criteria are calculated as mentioned in code (1). Then, the confidence score is calculated and applied to the sort algorithm as shown in code (2). The adopted base learner corresponds to the top of confidence score.

```
/***** Calculate confidence score *****/
int *confidence_score(int *SC)
{
    for (i=1;i<len;i++)
        CS[i]= SC[G1][i]/(SC[G2][i]+SC[G3][i]);
return CS;
}
/***** Sort confidence score *****/
int *sort_CS (int *CS)
{
    CS_Max=CS[sort(CS)];
    Pos_xy=*Pos(CS_Max);
return Pos_xy;
}
```

Code 2: Confidence score and sort functions

- Part 2: The decision is based in the lower hamming distance. We calculate the Hamming distance between the training data and the testing sample as defined in code (3).

```
/***** Calculate Hamming Distance *****/
int *Hamming_distance(int *TD, int *TS)
{
    for (i=vli;i<vli+5;i++)
        for (j=vci;i<vci+5;i++)
            Dist_TD_TS[i]=
Hamming(TD[i][j],TS[i][j]);
return Dist_TD_TS;
}
```

Code 3: Hamming distance function

In our case, the training data can be presented into three different pattern classes: (1) termination, (2) bifurcation and (3) non-Minutiae.

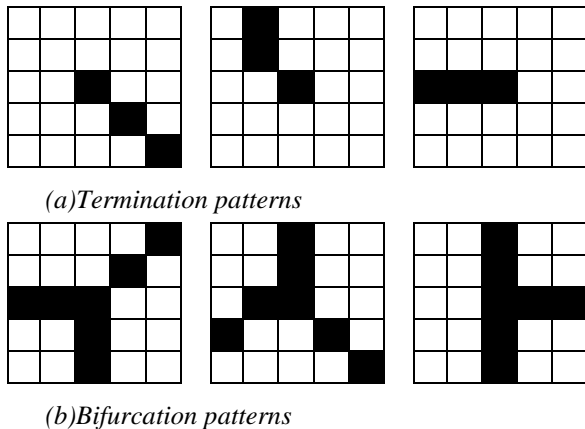


Fig. 4. Sample of different patterns

The used size mask for image is the  $5 \times 5$  pixels; it represents more information than the  $3 \times 3$  mask and sufficient information than the  $7 \times 7$  pixel mask. It is necessary that the size of the mask be an odd number to ensure the presence of central pixel.

To Apply the DECOC classifier, we have created 136 different patterns for two classes. 32 constructed patterns for termination class, 104 patterns for bifurcation class. Each

pattern sizes  $5 \times 5$  pixel. The third class, non-Minutiae, is detected when an input pattern is not considered as termination class or as bifurcation class. Figure 4 presents some constructed patterns for the class Termination (a) and for the class Bifurcation (b).

#### IV. THE MINUTIAE MATCHING PHASE

Matching phase can be presented as a mechanism that is responsible to make decision based on likeness parameter between two fingerprints. There are different methods in literature that lets Minutiae matching based on types of features [29][30][31] or on correlations of image [28].

The Minutiae matching phase aims (1) to make a reliable decision and (2) to ensure the real time response.

The first is faced with the problems of variations as the displacement of fingers, the rotation of fingers, the nonlinear distortion, the pressure of fingers when touch the sensor, the skin condition and feature extraction errors, etc [1], present a major problem.

Minutiae matching are classified in two families [28]:

- Global Minutiae matching; it aims to align Minutiae between two compared fingerprinted based on two directions and the angle. Works based on global Minutiae still suffer from on-precision and false identification [28].
- Local Minutiae matching; it aims to compare two fingerprints according to the relationships between proximity Minutiae. Works based on local Minutiae matching is more adopted than global Minutiae matching [28].

In the literature, there mainly three proposed method for matching phase. The first represents the classical matching algorithm [53]. In this case each Minutiae is identified by its neighborhood Minutiae and the comparison is made by pairs. When a two pairs is detected as similar, the rest of comparison is made by coordinates and angles.

The second Minutiae matching algorithm [54] considers the topology of Minutiae given by fixed radius and the comparison is ensured by the similarity of local topology.

The third Minutiae matching algorithm [55] gather the methods in [53] and [54]. Each Minutiae is described using its position and the relative position of neighboring Minutiae.

To sum up, the last method is more precise but its time consumption is important.

This paper proposes a modified algorithm based in [55]. The proposed matching algorithm uses the location coordinates to extract new information/ relationship between Minutiae without be tied with fixed reference, as shown in figure 5. In our case, Minutiae is presented by equation (7).

$$M = (x, y, \theta) \quad (7)$$

Where (x,y) is the coordinate of Minutiae

$\theta$  is the orientation angles



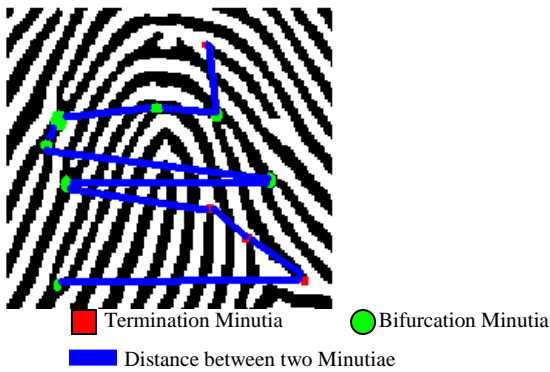


Fig. 5. The matching method

The matching method is based on three parameters:

- The Normalized Euclidean Distance between every two Minutiae is calculated according to equation (8),

$$Distance_{(M1,M2)} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (8)$$

where M1 and M2 two Minutiae and (x,y) presents the coordinate of each point.

- Direction between two successive Minutiae, see figure 6, using equation (9), (10),

$$Direction_{(M1,M2)} = \frac{y_2 - y_1}{x_2 - x_1} \quad (9)$$

$$\theta = |\arctan(direction1)| + |\arctan(direction2)| \quad (10)$$

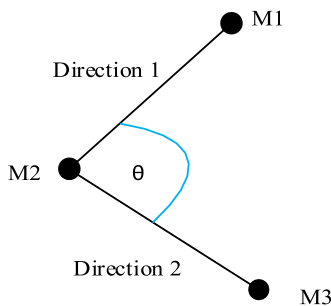


Fig. 6. Angle between three Minutiae

- Types of successive Minutiae as mentioned in equation (11),

$$Type_{(M1,M2)} = \begin{cases} 11 & \text{Bifurcation - Bifurcation} \\ 10 & \text{Bifurcation - Termination} \\ 01 & \text{Termination - Bifurcation} \\ 00 & \text{Termination - Termination} \end{cases}$$

At conclusion, every fingerprint is associated with a unique signature that is composed by distance, type and direction between two Minutiae, see equation (12).

$$M = (Distance_{MM-1}, Type_{MM-1}, \theta) \quad (12)$$

The matching algorithm process as follow:

- 1) Compare between Input Distance and the all stored distance in database. If the distance is lower than  $\epsilon=0.01$ , we proceed to next step.
- 2) Compare between Input Type and the adequate stored type in database. If the comparison is equal we proceed to next step.
- 3) Compare between Input direction and the adequate stored direction in database. If the direction is lower than  $\epsilon=0.05$  the Minutiae is accepted and it be considered a true Minutiae.

The second aims of the Minutiae matching phase in this paper, as mentioned above, is to ensure the real time execution. Many system based on identification is challenged by the real time decision and there are many related works in the literature that deals about it [42], [43], [44]. The real time execution is ensured by the identification for distributed database or by the time of identification.

In [50], the authors propose Minutiae matching phase based on client-server system when the fingerprint database is presented in many servers. But this kind of system doesn't support the parallel execution. That's why client-server system is not suitable for very large database.

In [51] and [52], the authors propose Minutiae matching phase based on agent-based systems. In fact this kind of system is adopted for the heterogeneous hardware architecture. The objective is to ensure load balancing between shared machines. Unfortunately, as mentioned in [51], the execution time is still important to guarantee the real time decision for large database.

The cited research works goals to provide identification for distributed databases, but this paper focuses to decrease the time of identification for large database. The state of arts is introduced in the next section.

This paper proposes, in the next section, an architecture based on GPU and multi-processor design for fingerprint recognition to accelerate the identification time.

## V. EXPERIMENT RESULTS

The implementation of fingerprint recognition represents a challenge especially for system that contains a huge number of fingerprints saved in database. Nowadays, the progress in term of technology became an advantage for recognition system to accelerate the consuming time. Fingerprint recognition involves many multiplications, evolution, rotation, and floating-point operations, which greatly slow the processing speed [24]. The implementation of fingerprint recognition system using hardware accelerator is widely used in literature [22] [23]. These implementations are much linked with the speed of the used processor and the time execution of the whole system is limited. In the next section, a brief resume for different hardware architectures is described.

- Hardware board; there are many work that implement the whole recognition algorithm into board based on microcontroller as [38]. But this solution suffers from the increasing of execution time above all when the database exceeds about one thousand.

- Embedded systems; this solution is adopted for intelligent sensors [39]. In this case the sensor encloses the whole architecture to make decision about recognition. Embedded systems are used for a limited number of users.
- Hardware acceleration fingerprint; this solution is performed to speed up filtering and matching step by using: FPGA-based [40], GPU-based [41] [43], parallel architecture [42].

In this paper, we propose to combine into hardware accelerator and multi-processor architecture. The next section presents the multi-processor architecture.

#### A. Multi-Processor architecture

The image of fingerprint is received by sensor with size of 200x300 pixels and divided into mask sized by 10x10 pixels. The proposed design is composed by SRAM memory, control unit, 4 processors elements (cores) and 4 hardware accelerators based on Gabor filter, see figure 7. The proposed architecture belongs to SIMD/MPSOC field [44]. The control unit represents an essential component in design. The last aims to handle all processes:

- Arbitrating the access processors units to/from memory;
- Handling and commanding all the processor unit;

At first, the image is captured with the fingerprint sensor and saved into principal memory. The proposed algorithm for fingerprint recognition is divided into software modules (skeletonization and Minutiae extraction) and hardware components (Filtering, Binarization and Matching), see the next section. In this case, the skeletonization and the extraction Minutiae code are saved, also, into principal memory. The control unit assigns the right process for processors elements. The proposed MPSOC architecture is composed by Microblaze soft-core processors interfaced with shared memory using the AXI4 bus [45]. All processors elements, on-chip memories, and the AXI4 bus are clocked at 100 MHz. Off-chip memory is clocked at 200 MHz and the AXI-lite bus is clocked at 50 MHz.

#### B. HW/SW implementation

This section presents the implementation of the recognition algorithm using Co-design approach. The aim of this section is to describe a hardware accelerator based on Gabor filter.

##### Step 1: Partitioning

Partitioning phase is based on execution time of different module of fingerprint recognition as shown in figure 8. The rule of partition suppose that the module where spend more time will be considered as hardware component and the module where spend less time will be considered as software parts [18]. Then, we find the result shown in figure 8 based on the native execution of the fingerprint recognition on a 6 Giga Byte memory, 2.2 GHz frequency of the processor Intel Core i3 with Windows 7 as operating system. We notice that the time execution of the Minutiae extraction module is the minimum and we can divide our architecture into:

- Hardware components: Binarization, Filter and Matching.
- Software applications: Minutiae extraction and Skeletonization.

##### Step 2: Gabor filter design

Gabor filter can be divided into three essential parts: Control Unit (CU), Arithmetic Unit (AU) and Memory [11]. The proposed architecture is based on 'CONV' signal that precede the convolution operation of the filter. When the 'CONV' signal is selected, the convolution process starts. When it is not selected, the filter proceeds in reception phase. In this phase the filter receives data that is corresponding on image input from DATAIN. The storage will be in the memory, see figure 9.

Simulation shows that the step of filtering consumes more time in execution than other steps. The idea is to implement a digital Gabor filter as a hardware accelerator. The contribution is to describe the Gabor phase that was never exploited [9] in the literature. There are three essential parameters that must be considered under design: (1) accelerate the speed of execution (2) minimize the Silicium area and (3) decrease the power consumption. Firstly, when the convolution signal is not selected the input data receives pixels and stores them in the memory [11]. When the convolution signal is selected the convolution process starts and follows steps mention in figure 10. The output image that is filtered is given after nine convolution operation.

Figure 11 presents the overview of the top level filter. Gabor filter is composed by six inputs and one output. Data IN receives initial data before filtering. During the writing phase in memory, P-X and P-Y will be stored in specific location. CLK input generates signal every 40ns period. RST input resets the execution of the filter. Arithmetic Unit presents the core of the digital Gabor filter, see figure 11. This unit store the coefficient of Gabor [4]. AU is composed by necessary 2 parts: the ROM and the MAC.

- The main advantage of the ROM as a memory is to save the coefficient.
- The MAC is the association of nine multipliers and eight adders. The multipliers are designed in parallel to accelerate the convolution process. The adders are designed in sequence to make the sum of the nine multipliers.

The 'CONV' signal ensures that the convolution process has made without any mismatch and with the correct data. The 'CLK' signal and the 'CONV' signal are connected to the memory and the MAC. When the CU receives 'READY' and 'SET', the convolution process finishes. The CU sends an activation signal to the 'CONV' to activate the convolution process. The convolution function takes 9 complete processes. Before the multiplier operation, the 'CON' signal data will be stored in a buffer. The idea is to be sure that the data sent from the memory to the convolution process is correct. The design is based on pipeline architecture. The duration of execution contains 222 clock cycles with a time period of 40ns.

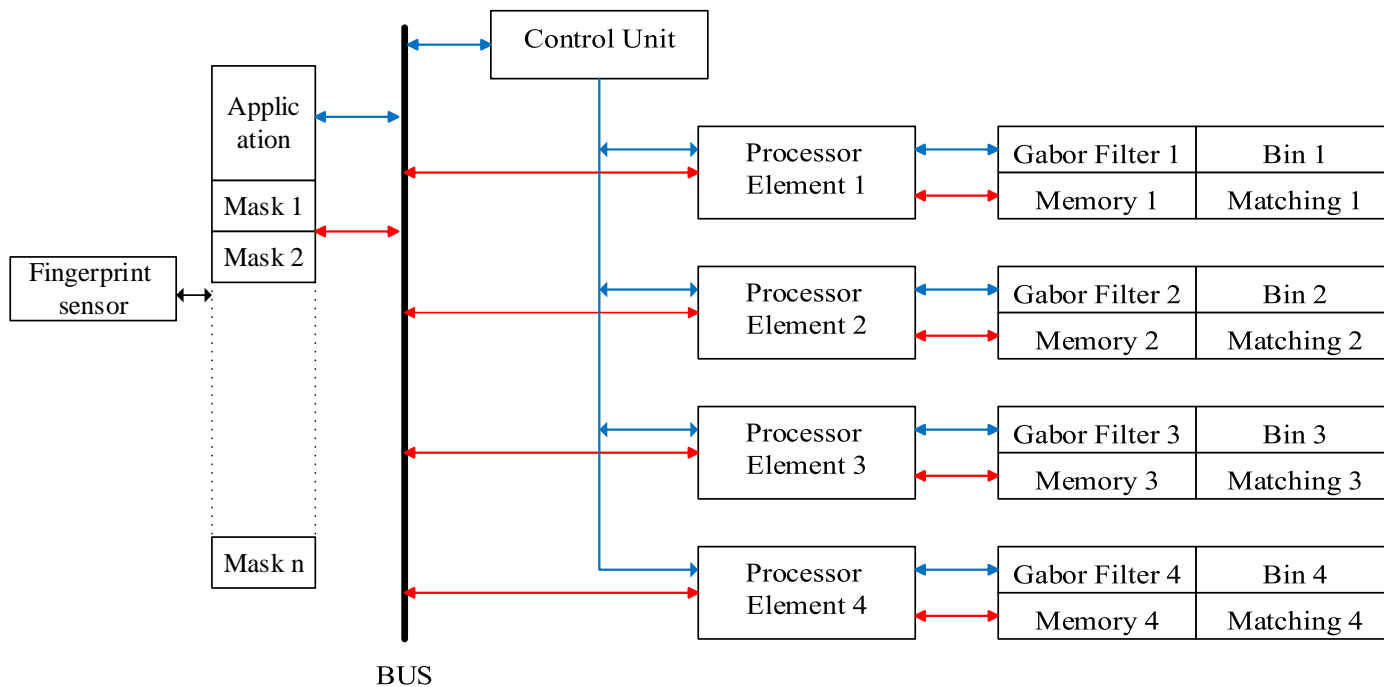


Fig. 7. Proposed MPSOC architecture

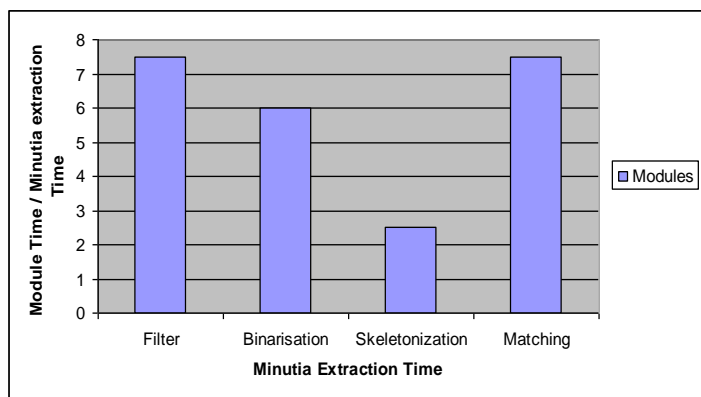


Fig. 8. Time execution of each modules as a function of Minutia extraction time

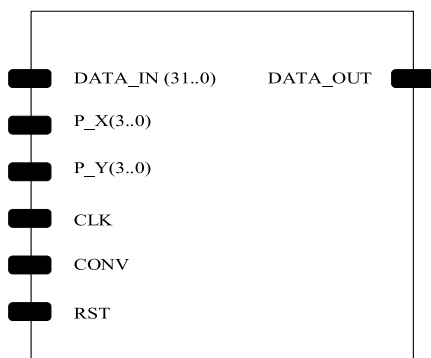


Fig. 9. Top level of Gabor filter

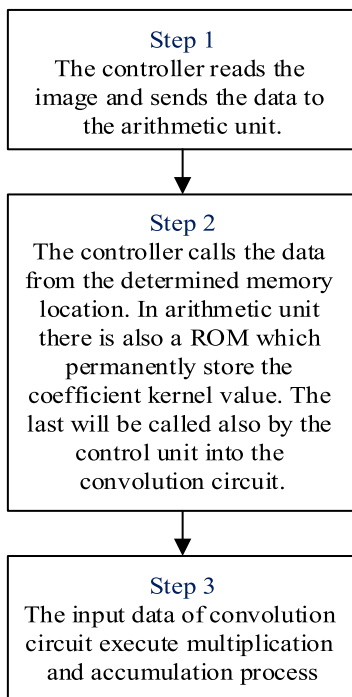


Fig. 10. Process Steps

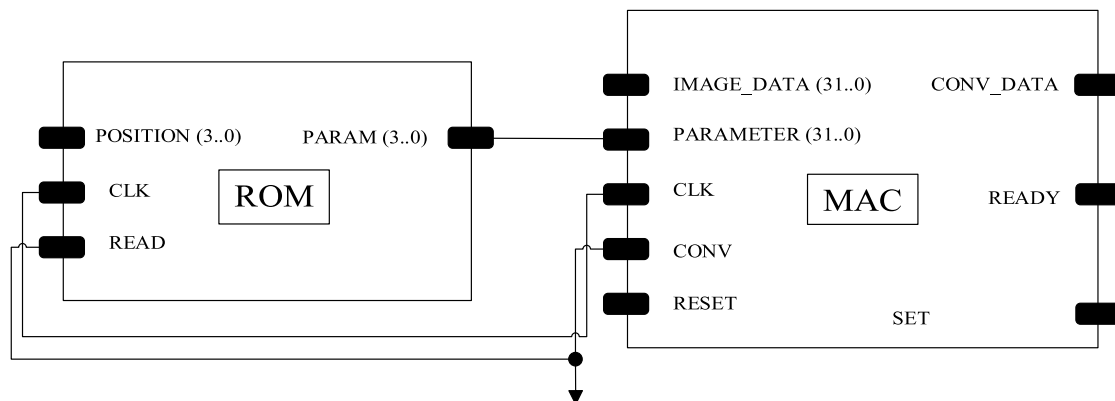


Fig. 11. Arithmetic unit

The control logic unit function presents the main controller in the filter. His role is to attribute process for other blocks. It plays also the role of middle between memory and AU.

The CU generates the address location if the 'START' signal is selected. The CU is composed by two blocks: counter for the coefficient and the counter decoder. The CU is composed by two blocks: counter for the coefficient and counter for the decoder.

The design of the counter gives the relationship between the coefficient and memory address.

The main role of the memory is to save data image in specific location. The address for X-direction is complete from two parts: CU or DATAIN.

C. Results and discussion

In this section, the results obtained from experiments are presented. Discussion section can be divided into two parts:

GPU (Gabor filter, binarization and matching) experiments, multi-processor design experiments.

**GPU (Gabor filter, binarization and matching) experiments;**

The outputs results for the top level filter are the real convoluted data and the expected result. The difference between outputs results represents the error of convolution function. This error is only about 0.001%. The proposed design of Gabor filter based on amplitude and phase both shows a good performance. In addition, the execution time is speedy, it takes 222 cycles only to obtain filtered image. In term of silicium surface, experiments show also that the number of slices has been decreased from 5759 to 1625. This result is performed because in our proposed design we have reduce the number of multipliers and adders compare of the traditional architecture of Gabor filter. Our proposed design makes different optimization in memory functionality and the

controller unit.

### Multi-processor design experiments;

DB14 represents a database that contains 10,000 fingerprints as templates. By a single identification, the recognition process computes 100 identifications using 100 different and randomly selected fingerprints from each database as the input [43]. The next table presents the results using the proposed algorithm and design comparing with others works. It is necessary to mention that all compared

works use the same database and the same conditions [43].

In table 2, the proposed design ensures the best execution time for fingerprint recognition. Our proposition shows that the speed-up factors range is about 44.63 times in comparison with CPU execution in [43]. The result shows also that the speed-up factors range is about 2.84 times in comparison with GPU execution in [43], 1.84 times in comparison with Jiang design [46], and 1.43 times in comparison with Chen design [47], see figure 12.

TABLE II. COMPARISON BETWEEN THE PROPOSED DESIGN AND RELATED WORKS BASED ON TIME EXECUTION

	Execution Time (s)	Speed-up in comparison with CPU [43]
Execution based on CPU [43]	10.802	
Execution based on GPU [43]	0.688	15.69
Jiang Design [46]	0.446	24.21
Chen Design [47]	0.347	31.12
<b>Our proposition</b>	<b>0.242</b>	<b>44.63</b>

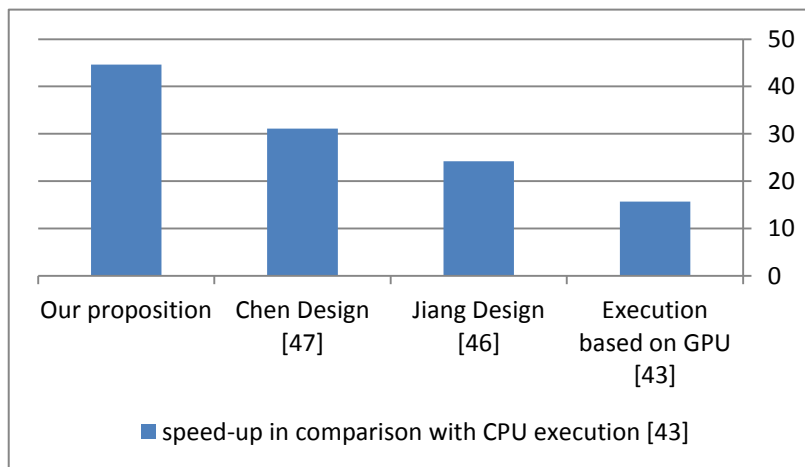


Fig. 12. speed-up factors range

### VI. CONCLUSION

In this paper, an algorithm for fingerprints recognition is proposed. The last is based on comparison made between different works in literature. As a first contribution, a Gabor filter design has been proposed as hardware accelerator to decrease the execution time. The estimated results are reached. The experiment results between the outputs results of Gabor filter algorithm and the design is too near. This proves that the proposed blocks are well chosen.

As a second contribution, a whole hardware design based on multi-processor architecture is described. The proposed design combines with 4-cores and GPU (Gabor filter, binarization and matching) for hardware implementation and software modules: skeletonization and extraction Minutiae.

Compared with other works cited in previous section, our design achieves the best execution time. This advantage is showed mainly for ATM system that has a huge number of saved fingerprints.

### ACKNOWLEDGMENT

The author would like to thank Deanship of Scientific Research at Majmaah University for funding this work.

### REFERENCES

- [1] d. Maltoni, d. Maio, a.k. jain, and s. Prabhakar, "handbook of fingerprint recognition". Springer; 1st edition, 2003, isbn 0- 387-95431-7.
- [2] b.g. sherlock, d.m. monro, and k. Millard, "fingerprint enhancement by directional fourier filtering," vision, image and signal processing, iee proceedings-, vol. 141, no. 2, 1994, pp. 87-94.
- [3] p.m. patil; s.r. suralkar; f.b. sheikh, "rotation invariant thinning algorithm to detect ridge bifurcations for fingerprint identification", 17th iee international conference ictai 05 on volume, issue, 16-16 nov. 2005.
- [4] n. Ratha, s. Chen, and a.k. jain, "adaptive flow orientation based feature extraction in fingerprint images," elsevier pattern recognition issn 0031-3203 coden ptnra8, vol. 28, no. 11, 1995, pp. 1657-1672.
- [5] a.farina, k. Vajna, and a.leone, "fingerprint minutiae extraction from skeletonised binary images," pattern recognition, vol. 32, no. 5, 1999, pp. 877-889.
- [6] d. Maio and d. Maltoni, "direct gray-scale minutiae detection in fingerprints," iee transactions on pattern analysis & machine intelligence, vol. 19, no. 1, 1997, pp. 27-41.
- [7] x. Jiang, w.y. yau, and w. Ser, "detecting the fingerprint minutiae by adaptive tracing the gray-level ridge", pattern recognition, vol. 34, no. 5, 2001, pp. 999-1013.
- [8] m. Ben ayed, f. Bouchhima and m. Abid, "automated fingerprint recognition using the decoc classifier", international journal of computer information systems and industrial management applications. Issn 2150-7988 volume 4 (2012) pp. 546-553
- [9] f.bellakhddhar , m.bousselmi and m. Abid, "face identification using the magnitude and the phase of gabor wavelets and pca", 2012

- [10] m. Liwen, y. Liang, q. Pan, h. Zhang "a gabor filter based fingerprint enhancement algorithm in wavelet domain", proceeding of iscit 2005.
- [11] p. H. W. L. Ocean y. H. Cheung, eric k.c. tsang, "implementing of gabor-type filters on field programmable gate arrays," 2005.
- [12] b. Heisele, p. Ho, "face recognition: component-based versus global approaches". *Computer vision and image understanding*, 2008.
- [13] w. Kong, d. Zhang, w. Li, "palmprint feature extraction using 2-d gabor filters", the journal of the pattern recognition society, elsevier, 2003, p 2339-2347
- [14] m. Abadi, m. Khoudeir, s. Marchand, "gabor filter-based texture features to archaeological ceramic materials characterization", *image and signal processing*, springer, 2012, volume 7340, pp 333-342
- [15] s. Shaikh, a. Maiti, n. Chaki, "a new image binarization method using iterative partitioning", *machine vision and applications*, 2013, volume 24, pp 337-350.
- [16] j. Teubner, l. Woods, c. Nie, "skeleton automata for fpgas: reconfiguring without reconstructing", *proceedings of the 2012 acm sigmod international conference on management of data*, 2012, pp 229-240.
- [17] a. Beristain, m. Grana, "a stable skeletonization for tabletop gesture recognition", *international conference on computational science and its applications*, 2010, pp 610-621.
- [18] l. Luo, h. He, q. Dou, w. Xu, "hardware/software partitioning for heterogeneous multicore soc using genetic algorithm" *international conference on intelligent system design and engineering application*, 2012, pp 1267-1270.
- [19] j. Silva, k. Costa, v. Roda, "the c compiler generating a source file in vhdl for a dynamic dataflow machine", *international conference on mathematical methods and computational in electrical engineering*, 2009, pp 33-36.
- [20] pradeepta k. Sarangi, p. Ahmed, kiran k. Ravulakollu, "naïve bayes classifier with lu factorization for recognition of handwritten odia numerals", *international journal of science and technology*, volume 7, issue 1, january 2014.
- [21] m. Kocevar, s. Klampfer, a. Chowdhury, z. Kacic, "low-quality fingerprint image enhancement on the basis of oriented diffusion and ridge compensation", *international journal elektronika ir elektrotehnika*, volume 20, issue 8, 2014.
- [22] m. Gok, s. Gorgunoglu, i. Muharrem orak, "fingerprint pre-processing on arm and dsp platforms", *international journal elektronika ir elektrotehnika*, volume 20, issue 6, 2014.
- [23] f. Yucel, o. Oral, n. Caglayan, m. Tecimen, s. Kocak, e. Yuce, "design and implementation of a personal computer authorization system using color detection", volume 15, issue 9, 2011.
- [24] j. Wang, l. Wu, y. Liu, "nios ii processor-based fingerprint identification system", *nios ii embedded processor design contest—outstanding designs 2007*.
- [25] a t gowthami, h r mamatha, "fingerprint recognition using zone based linear binary patterns", *procedia computer science*, v.58, 552-557, 2015.
- [26] s. Prabhakar, anil k. Jain and sharath pankanti. *Learning fingerprint minutiae location and type*. pattern recognition society. Published by elsevier science ltd, 2003 .
- [27] m. Ben ayed, f. Bouchhima, m. Abid, "a novel application of the classifier decoc based on fingerprint identification", *interactive multimodal pattern recognition in embedded systems impress 2010 workshop on database and expert systems applications dexa 2010*.
- [28] d. Peralta, m. Galar, i. Triguero, daniel paternain, salvador garcia, edurne barrenechea, josé m. Benitez, humberto bustince, francisco herrera, "a survey on fingerprint minutiae-based local matching for verification and identification: taxonomy and experimental evaluation", *information sciences journal*, p 67-87, 2015
- [29] a.k. jain, j. Feng, k. Nandakumar, fingerprint matching, *iee comput*. 43 (2010) 36-44.
- [30] d. Maltoni, d. Maio, a.k. jain, s. Prabhakar, *handbook of fingerprint recognition*, second ed., springer publishing company, incorporated, 2009.
- [31] n. Yager, a. Amin, fingerprint verification based on minutiae features: a review, *pattern anal. Appl*. 7 (2004) 94-113.
- [32] h. Guo "a hidden markov model fingerprint matching approach", *proceedings of the fourth international conference on machine learning and cybernetics*, guangzhou, iee print isbn: 0-7803-9091-1 , 5055 - 5059 vol. 8 , august 2005
- [33] o. Saeed, a. Bin mansoor, and m. Asif afzal butt "a novel contourlet based online fingerprint identification", *bioid multicommm2009*, lncs 5707, pp. 308-317, 2009. Springer-verlag berlin heidelberg 2009.
- [34] y. Hao, t. Tan, y. Wang "an effective algorithm for fingerprint matching" *national lab of pattern recognition*, cas, institute of automation, beijing, p. R. China.
- [35] j. Zang, j. Yuan, f. Shi, s. Du "a fingerprint matching algorithm of minutia based on local characteristic", isbn 978-0-7695-3304-9/08, iee 2008.
- [36] ch.tang hsieh and ch. Hu, "fingerprint recognition by multi-objective optimization pso hybrid with svm", *journal of applied research and technology*, volume 12, issue 6, december 2014, pages 1014-1024
- [37] j. Hong, j. Min, u. Cho, s. Cho, "fingerprint classification using one-vs-all support vector machines dynamically ordered with naïve bayes classifiers ", *pattern recognition*, volume 41, issue 2, february 2008, pages 662-671
- [38] m.a. murillo-escobar, c. Cruz-hernandez, f. Abundiz-pérez, r.m. lopez-gutiérrez, "a robust embedded biometric authentication system based on fingerprint and chaotic encryption", *expert systems with applications*, volume 42 (2015) 8198-8211
- [39] s. Bayram, h.t. sencar, n. Memon, efficient sensor fingerprint matching through fingerprint binarization, *iee trans. Inform. Forensic. Secur.* 7 (2012) 1404-1413.
- [40] r.m. jiang, d. Crookes, fpga-based minutia matching for biometric fingerprint image database retrieval, *J. Real-time image process.* 3 (2008) 177-182.
- [41] p.d. gutiérrez, m. Lastra, f. Herrera, j.m. benitez, "a high performance fingerprint matching system for large databases based on gpu", *iee trans. Inform. Forensic. Secur.* 9 (2014) 62-71.
- [42] d. Peralta, i. Triguero, r. Sanchez-reillo, f. Herrera, j.m. benitez, "fast fingerprint identification for large databases", *pattern recogn.* 47 (2014) 588-602.
- [43] m. Lastra, j. Carabaño, p. D. Gutiérrez, j. M. Benítez, f. Herrera, "fast fingerprint identification using gpus ", *information sciences*, volume 301, 20 april 2015, pages 195-214
- [44] d. Watson, a. Ahmadinia, " memory customisations for image processing applications targeting mpsocs ", *integration, the vlsi journal*, volume 51, september 2015, pages 72-80
- [45] s. Saadi, m. Touiza, f. Kharfi, a. Guessoum, " dyadic wavelet for image coding implementation on a xilinx microblaze processor: application to neutron radiography ", *applied radiation and isotopes*, volume 82, december 2013, pages 200-210
- [46] x. Jiang, w. yau, " fingerprint minutiae matching based on the local and global structures", *proceedings of the 15 th international conference on pattern recognition*, vol.2, iee, 2000, pp. 1038-1041.
- [47] k. Cao, x. yang, x. chen, y. zang, j. liang, j. tian, "a novel ant colony optimization algorithm for large-distorted fingerprint matching", *pattern recognition* 45(2012)151-161.
- [48] r. Cappelli, m. ferrara, d. maltoni, "minutia cylinder-code: a new representation and matching technique for fingerprint recognition", *iee transactions on pattern analysis and machine intelligence* 32(2010)2128-2141.
- [49] nemeth, g., k. Palagyi., topology preserving parallel thinning algorithm. *International journal of imaging system and technology*, 2011: 21: 37-44.
- [50] r.f.miron, t.s. letia, m. hulea, "two server topologies for a distributed fingerprint-based recognition system" ,15th international conference on system theory, control and computing, 2011, pp.1-6.
- [51] k. Beghdad bey, z. Guessoum, a. mokhtari, f. benhamadi, "agent based approach for distribution of fingerprint matching in a meta computing environment", *proceedings of the 8th international conference on new technologies in distributed systems*, 2008, pp.1-7.



- [52] k. Nagaty, e.hattab, "an approach to a fingerprints multi-agent parallel matching system", iee international conference on systems, man and cybernetics, volume5, 2004 ,pp.4750–4756.
- [53] x. Jiang, w.yau , "fingerprint minutiae matching based on the local and global structures", proceedings of the 15th international conference on pattern recognition, vol.2, ieee, 2000, pp.1038–1041.
- [54] x. Chen, j.tian, x.yang, "a new algorithm for distorted fingerprints matching based on normalized fuzzy similarity measure", ieee transactions on image processing 15(2006)767–776.
- [55] r. Cappelli, m.ferrara, d.maltoni, "minutia cylinder-code: a new representation and matching technique for fingerprint recognition", ieee transactions on pattern analysis and machine intelligence (2010)2128–2141.

# Planning And Allocation of Tasks in a Multiprocessor System as a Multi-Objective Problem and its Resolution Using Evolutionary Programming\*

Apolinar Velarde Martínez  
Department of Computer Science  
Technological Institute El Llano Aguascalientes  
El Llano, Aguascalientes México

Juan Antonio Nungaray Ornelas  
Department of Computer Science  
Technological Institute El Llano Aguascalientes  
El Llano, Aguascalientes México

Eunice Ponce de León Sentí  
Department of Computer Science  
Universidad Autónoma de Aguascalientes  
Aguascalientes, México

Juan Alejandro Montañez de la Torre  
Department of Computer Science  
Technological Institute El Llano Aguascalientes  
El Llano, Aguascalientes México

**Abstract**—the use of Linux-based clusters is a strategy for the development of multiprocessor systems. These types of systems face the problem of efficiently executing the planning and allocation of tasks, for the efficient use of its resources. This paper addresses this as a multi-objective problem, carrying out an analysis of the objectives that are opposed during the planning of the tasks, which are waiting in the queue, before assigning tasks to processors. For this, we propose a method that avoids strategies such as those that use genetic operators, exhaustive searches of contiguous free processors on the target system, and the use of the strict allocation policy: First Come First Serve (FCFS). Instead, we use estimation and simulation of the joint probability distribution as a mechanism of evolution, for obtaining assignments of a set of tasks, which are selected from the waiting queue through the planning policy Random-Order-of-Service (ROS). A set of conducted experiments that compare the results of the FIFO allocation policy, with the results of the proposed method show better results in the criteria of: utilization, throughput, mean turnaround time, waiting time and the total execution time, when system loads are significantly increased.

**Keywords**—Multicomputer system; Evolutionary Multi-objective Optimization; First Input First Output; Random-Order-of-Service; Estimation of Distribution Algorithms; Univariate Distribution Algorithm

## I. INTRODUCTION

Multi computer systems with architectures and mesh topologies using 2D and 3D interfaces, designed for commercial and research purposes, have been two of the most common networks in research and industrial environments because of their simplicity, scalability, structural regularity and ease of implementation [1, 2, 3]. Examples of such systems are the IBM BlueGene / L [4] and the Intel Paragon [5]. Some of the commercial Multi Computer systems are Multiple Instruction Multiple Data (MIMD) systems with architectures that enable partitions of processor submeshes, and have the advantage of supporting multiple parallel (multi-tasks ) jobs [1, 2, 3, 6]. Parallel jobs are usually represented by a Directed Acyclic Graph (DAG), the nodes express the particular tasks

partitioned from an application and the edges represent the inter-task communication [7]. The tasks can be dependent or independent; independent tasks, can be executed simultaneously to minimize processing time, and dependent tasks are cumbersome and must be processed in a pre-defined manner, to ensure that all dependencies are satisfied [6]. In an SIMD mesh, that processes parallel jobs, tasks are planned in the queue by a planning policy (usually being First Come First Serve (FCFS)) [2, 3, 8, 9], they are then assigned to the mesh processor, where they remain until they finish their implementation [7]. Planning of resources in the mesh, through hardware partitioning involves two components: a scheduler and dispatcher to the mesh [2, 3, 8, 9]. The function of the scheduler is to choose the next task, or the following tasks in the queue that will be assigned to a sub-mesh, of free processors for execution. The function of the submesh allocator is to locate free submeshes, which are to be assigned to the selected tasks by the scheduler. The allocator uses a contiguous and/or of noncontiguous assignment method. When a contiguous allocation method is used, the tasks partitioned from an application can only be assigned to adjacent processors, unlike a noncontiguous allocation method, where tasks can be assigned in a scattered form across the mesh wherever free processors are located [2,3,8,9]. To maximize the use of resources in the target system, current computer systems opt to use non-contiguous allocation methods, applying wormhole routing and free submesh recognition techniques.

Some examples of this are: the frames processor that uses windows traveling the length and width of the grid [10]; iterative processes that divide submeshes in equal-sized partitions [11], the use of the free-lists approach, [12, 13] among others [22-26]. During the processing of tasks extracted from the queue, we look to optimize a set of objectives that are generally found to be opposite. Upon finalizing the total processing of tasks running on the target system, we seek to optimize a set of proper criteria, from the multiprocessor systems. In the following paragraphs, we list the objectives and

This investigation is sponsored by Tecnológico Nacional de México, and developed in Instituto Tecnológico el Llano, Aguascalientes, México.

exemplify the form in which they are opposed, as well as the list of criteria that is optimized.

The objectives sought to be optimized, for processing tasks using the scheduler and proposed in [14] are:

1) *Reduce the waiting time of tasks in the queue, assigning more tasks to the mesh of processors once the allocator reports, the number of processors in the free submeshes.*

2) *Reduce task starvation, that is, avoid discrimination in the allocation of tasks that require a lot of processors (great tasks), caused by the continued allocation of tasks requiring a lesser amount of processors (small tasks).*

3) *Minimize external fragmentation, that is, minimize the percentage of free processors, after the allocation algorithm places one or more tasks in the processor mesh.*

4) *Minimize the communication overhead or network contention [15], through contiguity between processors (as close as possible to assign the set of free processors), in order to decrease the distance in the communication path, and avoid interference between the processing elements (searches for the best way to accommodate tasks in the free processors). This point is then identified as the quadratic dynamic allocation of tasks.*

Upon complete processing of all tasks in the system, we then seek to optimize the system utilization criteria, throughput, response rate, mean turnaround time and overall waiting time [15, 16].

A simple example of the contrast of the previous four objectives occurs, when we look to minimize external fragmentation by using a noncontiguous allocation method. The largest number of jobs in the free processors is allocated regardless of their location in the mesh, resulting in the maximization of system utilization; however, the opposite effect is produced upon maximizing the communication overhead, between tasks if not assigned contiguously along the length and width of the grid. Thus, in seeking to maximize or minimize some of the objectives, in order to optimize your results, usually the result of another objective is degraded, producing contrasting results between themselves, enabling oneself to view the problem of task planning and allocation as a multi-objective problem.

A multi-objective problem, involves optimizing a number of targets simultaneously, and its solution with or without the presence of constraints, results in a set of interchangeable optimal solutions called the search space, popularly known as Pareto-optimal solutions. For an adequate solution in this, evolutionary optimization algorithms are utilized (EOA), which use a population focuses in their search procedure [17]. The EOA's possess several characteristics that are desirable for problems, involving multiple conflicting objectives and intractably large and highly complex search spaces [18].

In this paper, a hybrid method is proposed to address the problem of planning and allocation of multiple parallel jobs in a multiprocessor system, as a multi-objective problem. In this manner, it makes use of the scheduler and allocator to achieve the best assignments in the processor grid, which optimize the

resources of the target system, during processing and completion of tasks. This method uses a static task scheduling, defined as a scheduling at compile time [19].

The proposed method is evaluated with two task selection policies from the queue: FIFO and ROS [20]. This method connects the planner and the dispatcher to conduct the process of task selection, from the queue randomly and makes the best assignment in the processor grid, by evaluating a set of conflicting objectives. The work that the scheduler and dispatcher does is divided into five steps as follows: first, the dispatcher reports the number of free processors that the grid has in time  $t$ , in the second step, by means of the Random-Order policy-of-service the scheduler selects the same number of tasks, with subtasks from the queue that the allocator previously reported, regardless of the location of the processors across the grid. This set of selected tasks, is considered a feasible solution of the search space, to which three disjointed objectives are evaluated: the waiting time from the rest of the tasks that remain in the queue, the starvation of the tasks of the queue (if occurs), the external fragmentation, and the communication overhead. In the third step, the process of dynamic selection of tasks by the planner continues until to the stop criterion is fulfilled. Lastly, for the set of feasible solutions, the joint probability distribution can be appreciated using the algorithm UMDA (Univariate Marginal Distribution Algorithm), to obtain the best allocation to the processor grid. After finalizing the total execution of tasks, the following criteria is evaluated and compared: system utilization, throughput, response rate, mean waiting time and turnaround time with different workloads in the target system; the effectiveness of the proposed method is compared with the most widely used task planning method: FCFS.

This paper is organized as follows: In section 2, we discuss a classification of methods that throughout the years have been proposed for the planning and allocation of tasks using heuristics techniques and geometric models. In section 3, a stopping criterion is performed using a definition of the objectives, and the form in which they are opposed during the execution of tasks. Section 4, describes the functionality proposed in this research method. Section 5, describes the experiments conducted by the method. In section 6, the future work to develop after this research is described, and section 7, conclusions, describes the findings of this research.

## II. RELATED WORKS

In [19], two classes or categories for scheduling are specified, a) list scheduling and b) clustering; in this paper, related jobs are classified depending on 1) planner use and heuristic techniques, and 2) allocator use in conjunction with geometric patterns, and free submesh searches throughout the grid.

### A. Task scheduling methods that make use of the planner and heuristic techniques

Heuristic methods base their functionality in genetic algorithms (GA), which are global search techniques that explore different regions of the search space simultaneously, by keeping track of sets of potential solutions called a population [21]. Over the years, different methods have been

proposed based on this search technique. In this section we can observe how a set of these investigations show the similarities in the operators used.

In [16], the multiprocessor scheduling problem is based on the deterministic model, and the precedence relationship among the tasks is represented by an acyclic directed graph. This method uses a representation based on the schedule of the tasks, in each individual processor. Several lists of computational tasks represent the planner, respecting the order of precedence of the tasks.

Each list can be further viewed as a specific permutation of the tasks in the list. Crossover operators, reproduction and mutation are applied to the created lists, in order to optimize the finishing time of the schedule. Similar research is presented in [22], where the scheduling problem is formulated in a genetic search framework based on the observation, that if the tasks of a parallel program are arranged properly in a list, an optimal schedule may be obtained by scheduling the tasks, one by one according to their order in the list. These lists are codified in chromosomes, which represent feasible solutions in the search space; genetic search operators are applied to these chromosomes, such as crossover and mutation, as well as an additional operator called "an investment operator". Chromosomes are manipulated by genetic operators, in order to determine an optimal scheduling list, leading to an optimal schedule. To improve convergence time of the proposed algorithm, the connected synchronous island model is used. In [23], the proposed genetic algorithm minimizes the schedule length of a task graph, to be executed on a multiprocessor system. It uses processes that evolve candidate solutions by the use of a set of operators, such as fitness-proportionate reproduction, crossover and mutation; doing so through a traditional method of genetic algorithms. This algorithm does not consider the communication time between tasks. In [24], an initial chromosome, consisting of genes is generated, where each gene will use the priority of node in a directed acyclic task graph (DAG); trade, crossover and mutation operators are applied to the chromosome in order maximize the makespan of the  $k$ -th chromosome, using an evaluation function. Communication costs are not considered in this paper. In [25], a Modified List Scheduling Heuristic (MLSH) and hybrid approach composed of genetic algorithms, and MLSH for task scheduling in multiprocessor systems is used; this method uses three new different types of chromosomes: task list, list processor and a combination of both types. In order to maximize the finishing time of schedule, the genetic operators: crossover and mutation, used in the chromosomes are the selection. The main features of this type of method are the use of a set of genetic operators (parameters), which seek to optimize a single objective function (maximize execution times of the tasks). Nevertheless, if a researcher does not have experience in using this type of an approach for the resolution of a concrete optimization problem, then the choice of suitable values for the parameters can be converted into an optimization problem [26]. Similarly, in these methods as the complexity of the task graphs and the proposed solution are increased, the number of operators, that must manipulate the algorithm to try to find the best solutions in the search space also increases;

best case scenario being the possibility that the algorithm will land in the least amount of local minimums.

### B. Methods with geometric models for scheduling

Alternately to heuristic methods, other methods seek to solve the problem of task planning and allocation by looking for free processors, that are contiguous to the length and width of the grid; ensuring that the tasks assigned during implementation remain as close together as possible.

In [27], a submesh reservation strategy for incoming tasks is used, this method combines a submesh reservation technique with a priority technique as follows: an incoming task requests a number of processors, a reservation will occur if these cannot be assigned to of a set of processors constituted in a sub-mesh, as long as it does not exceed the threshold established within the parameter FREE\_FRAC. The priority of waiting tasks is handled through a "no\_supercede" parameter, which allows you to suspend allocations if the threshold in the parameter MAX\_PRI is exceeded, and it also prioritizes tasks that have aged in the waiting queue. In [28], the approach contains a list of allocated submeshes, sorted in a non-increasing order by the second coordinate in their upper right corner. This list serves two purposes: first, it determines the nodes that cannot be used as a basis for new free submesh applications and second, it identifies nodes that are located on the right edge of the assigned submeshes, in order find the nodes that could be used as a basis in finding free submeshes. When a parallel job is selected to be assigned, a search is performed to locate a suitable sub-mesh, if this does not occur; the assignment is made with longest free submesh, whose length of sides does not exceed the requested submesh. Through a search process, the free submesh that best fits the application is located. Other current techniques, through an initial strategy, look to make the allocation of tasks to the mesh, but if resulted in failure, a second allocation strategy is activated to replace the first in order achieve the assignment. For example in [29], the First Fit technique (FF) proposed in [30], that searches for free submeshes best suiting the application (to find the maximum adjacency between processors while reducing communication latency between tasks), is used in conjunction with the Best Fit (BF) technique proposed in [31]. This technique searches for the exact number of processors that the task requires in the free submeshes; thus, in [29], if a task requests a 4x4 sub-mesh and the request cannot be granted, the request size is reduced by a multiple of 2, then a 2x2 grid will be requested and so on until the request is the minimum number of processors, 1 X 1 in this case. When the first technique fails, the second technique BF is enabled, and through this, a search is performed within the free submeshes which best fit, that is, with the exact number of processors that the task requires [31].

## III. BASIC CONCEPTS

This section describes the concepts and the evolutionary algorithm used in this research.

### A. Definitions

Definition 1. An n-dimensional mesh has  $k_0 \times k_1 \times \dots \times k_{n-2} \times k_{n-1}$  nodes, where  $k_i$  is the number of nodes along the length of the  $i$ -th dimension and  $k_i \geq 2$ . Each node identified by  $n$  coordinates:  $\rho_0(a), \rho_1(a), \dots, \rho_{n-2}(a), \rho_{n-1}(a)$  where  $0 \leq \rho_i(a) < k_i$

for  $0 \leq i < n$ . Nodes  $a$  and  $b$  are neighbors if and only if  $\rho_i(a) = \rho_i(b)$  for all dimensions except for dimension  $j$ , where  $\rho_j(a) = \rho_j(b) \pm 1$ . Each node in a mesh refers to a processor and the two neighbors are connected by a direct communication link.

**Definition 2:** A 2D mesh, which is referenced as  $M(W, L)$  consists of  $W \times L$  processors, where  $W$  is the width of the mesh and  $L$  is the height of the mesh. Each processor is denoted by a pair of coordinates  $(x, y)$ , where:  $0 \leq x < W$  and  $0 \leq y < L$ . A processor is connected by a bidirectional communication link to each of its neighbors. For each 2D mesh  $2D = P_{ij}$ .

**Definition 3:** In a 2D mesh,  $M(W, L)$ , a sub-mesh:  $S(w, l)$  is a two-dimensional mesh belonging to  $M(W, L)$  with width  $w$  and height  $l$ , where  $0 < w \leq W$  and  $0 < l \leq L$ .  $S(w, l)$  are represented by the coordinates  $(x, y, x', y')$ , where  $(x, y)$  is the lower left corner of the submesh and  $(x', y')$  is the upper right corner. The node in the lower left corner is called the base node of the sub-mesh, and the upper right corner is the end node. In this case  $w = x' - x + 1$  and  $l = y' - y + 1$ . The size of  $S(w, l)$  is:  $w \times l$  processors.

**Definition 4:** In a 2D mesh  $M(W, L)$ , an available sub-mesh  $S(w, l)$  is a sub-mesh that meets the conditions:  $w \geq \alpha$   $l \geq \beta$  assuming that the required allocation of  $S(\alpha, \beta)$  refers to selecting a set of available processors for task arrival.

**Definition 5:** The correspondence of a task or subtask to a free processor in the mesh is defined as the following: if  $\mathcal{Q}$  is a set of system tasks, and  $\mathcal{Q} = J_1, J_2, \dots, J_n$  where  $n$  is the number of tasks in time  $t$  and  $\mathcal{Q}_k$  is a set of sub-tasks of task  $k$  where:  $\mathcal{Q}_k = j_{k1}, j_{k2}, \dots, j_{kf}(k)$  and  $f(k)$  is the total number of sub-tasks of task  $k$ . For each task  $j$  and each sub-task  $f(k) \in j$  has a processor in  $m_i \in P$  in which it is to execute each task  $j$  and each sub-task  $j_{kf}(k)$ , consuming an uninterrupted time  $t \in N$ .

**Definition 6:** Given two matrices size  $n \times n$ : a flow matrix  $F$  whose  $(i, j)$ -th elements represent flows between tasks  $i$  and  $j$  and an arrangement of distances  $D$ , whose  $(i, j)$ -th elements represent the distance between sites  $i$  and  $j$ . An assignment is represented by vector  $p$ , which is a permutation of the numbers  $1, 2, \dots, n$  and  $p(j)$  is where the task  $j$  is assigned. Thus, the quadratic task assignments can be written as:

$$\min_p \varepsilon \sum_{i=1}^n \sum_{j=1}^n f_{ij} dp(i)p(j) \tag{1}$$

**Definition 7:** An optimization problem is one whose solution involves finding a set of candidate alternative solutions that best meet the objectives. Formally, the problem consists of the solution space  $S$  and function objective  $f$ . Solving the optimization problem  $(S, f)$  consists of determining an optimal solution, namely, a feasible solution  $x^* \in S$  such that  $f(x^*) \leq f(x)$ , for any  $x \in S$ . Alternative solutions can be expressed by assigning values to some finite set of variables  $X = \{X_i; i = 1, 2, \dots, n\}$ . If  $U_i$  is denoted the domain or universe (set of possible values) of each of these  $n$  variables. The problem consists of selecting the value  $x_i$  that is assigned to each variable  $X_i$  from domain  $U_i$  that when subjected to certain restrictions, optimizes an objective function  $F$ . The universe of solutions is identified with the set  $U = \{x = (x_i; i=1, 2, \dots, n)\}$ .

$x_i \in U_i\}$ . The problem constraints reduce the universe of solutions to a subset of  $S \subseteq U$  called feasible space.

A performance evaluation of a parallel system, upon finalizing the processing of all running tasks, is evaluated on the following criteria [1]:

**Definition 8:** The utilization is defined as the fraction of time in which the system was used, and is given by:

$$U_G = W_G / (C_G * m_G) \tag{2}$$

Where:  $W_G$  is the amount of work that the system performs,  $C_G$  is the completion time of execution of all tasks in the system and  $m_G$  is the total number of processors in the system.

**Definition 9:** Throughput. The number of completed tasks per unit of time in the system is given by:

$$n / C_G \tag{3}$$

Where:  $n$  is the total number of jobs in the system.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

**Definition 10:** Mean turnaround time. The average time it takes all tasks from entering the local queue until their execution is finalized. Calculated as:

$$\frac{1}{n} \sum_{j=1}^n t_i^j \tag{4}$$

Where:

$$t_i^j = c^j - r^j$$

$c^j$  is the completion time of the task and  $r^j$  is the delivery time of task  $j$ .

**Definition 11:** Waiting time, defined as the average waiting time before starting the task execution. Calculated as:

$$\frac{1}{n} \sum_{j=1}^n t_w^j \tag{5}$$

Where:

$$t_w^j = t_s^j - r^j$$

$t_s^j$  is the start time of execution of task  $j$ .

**Definition 12:** Response ratio, defined as the coefficient response average of all tasks. Defined as:

$$\frac{1}{n} \sum_{j=1}^n (t_w^j + p^j) / p^j \tag{6}$$

Where:  $p^j$  is the runtime and  $t_w^j$  is the waiting time of task  $j$ .

**B. UMDA for dynamic quadratic assignment to model the problem of task scheduling**

The EDA (Distribution Evolutionary Algorithm) uses estimation and simulation, from the joint probability distribution as a mechanism of evolution, instead of, directly manipulating the individuals that represent solutions to the

problem [26]. An EDA begins by randomly generating a population of individuals, which represent solutions to the problem, iteratively performs three types of operations on the population: a subset of the best individuals of the population is generated; a learning process from a probability distribution model from selected individuals is performed, and new individuals that simulate the obtained distribution model are generated. The algorithm stops when a certain number of generations are reached, or when the performance of the population fails to significantly improve; an UMDA is used to estimate the joint distribution in each generation from selected individuals. Thus, the joint probability distribution is factorized as a product of independent univariate distributions, i.e.:

$$p_i(x_i) = \frac{\sum_{j=1}^N \delta_j(X_i = x_i | D_{i-1}^{S_i})}{N}$$

Where:

$$\delta_j(X_i = x_i | D_{i-1}^{S_i}) = \begin{cases} 1 & \text{si en el } j\text{-ésimo caso de } D_{i-1}^{S_i}, X_i = x_i \\ 0 & \text{en otro caso} \end{cases}$$

The pseudocode for an UMDA algorithm is as follows:

Generate  $M$  individuals (the initial population) randomly

Repeat for  $l = 1, 2, \dots$  until the stop criterion:

$D_{l-1}^{S_i} \leftarrow$  Select  $N \leq M$  individuals of  $D_{l-1}$  in accordance to the selection method

$$p_l(x) = \prod_{i=1}^n p_l(x_i) = \prod_{i=1}^n \frac{\sum_{j=1}^N \delta_j(X_i = x_i | D_{l-1}^{S_i})}{N}$$

Estimate the joint probability distribution.

$D_l \leftarrow$  Sample  $M$  individuals (the new population) from  $p_l(x)$

#### IV. STATEMENT OF THE PROPOSED METHOD

This section is structured as follows: Section 4.1 shows three instantiations of the manner in which the objectives are opposed during the planning and allocation of tasks. In section 4.2 the functionality of the proposed method is explained in detail.

##### A. Contraposition of the objectives during job processing

The way that the objectives are opposed during job processing is shown in [32], it is explained through three examples in the following sections. In Figure 1, an 8x8 2D processor mesh is shown; the 35 occupied processors are shown in closed circles and the 29 free processors, with unfilled circles. In the queue, a set of 6 dependent tasks partitioned from an application wait for execution: task  $T_0$  with 4 subtasks, task  $T_1$  with 3 subtasks, task  $T_2$  with 4 subtasks, task  $T_3$  with 3 subtasks, task  $T_4$  with two subtasks and task  $T_5$  with 25 subtasks, supposing that the planning method can choose more than one task to be assigned in the processor mesh with noncontiguous allocation method.

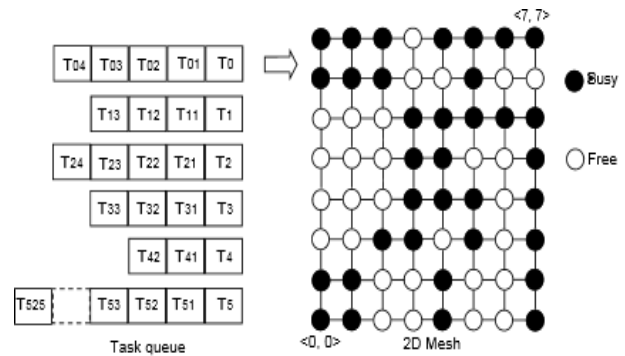


Fig. 1. System structure for task execution on a Multicomputer 2D mesh system

Example 1: Consider that in time  $t$ , the allocator reports the 29 free processors, with this data the scheduler determines that the set of 5 tasks:  $T_0$ ,  $T_1$ ,  $T_2$ ,  $T_3$  and  $T_4$  are candidates to occupy 21 processors in the mesh, or assign task  $T_5$  requiring 26 processors and task  $T_4$  requesting 3 processors.

By assigning the set of the 5 tasks, the same number of positions in the queue are released allowing the entry of new tasks, and the number of accesses to the queue is decreased in order perform more task searches. The previous procedure allows for more than one task to enter the mesh, and decreases task waiting time at the head of the queue; but in opposition to each other, the assigning of these 5 tasks generates an external fragmentation of 8 processors and produces starvation of task  $T_5$  in this assignment. The result that is had is a contrast between objectives 1 and 2.

Objective 1 seeks to minimize the number of assignments to the mesh of processors in order minimize task waiting time, and objective 2 seeks to maximize the use of the processors in the mesh and minimize starvation of the large tasks. Now, if tasks  $T_4$  and  $T_5$  are assigned, neither starvation nor external fragmentation occurs, but a smaller number of tasks can be accepted in the queue, and so the number of assignments to the mesh increases therefore, also increasing the time tasks must wait to enter the processor mesh.

Example 2: In order to illustrate the contrast between objectives 3 and 4, consider Figure 1. Objective 3 seeks to maximize the use of the processors in the mesh, avoiding external fragmentation, and Objective 4 seeks to minimize overhead communication through minimizing the adjacency of processors that are assigned to a task. The assumed set of the 5 selected tasks are:  $T_0$ ,  $T_1$ ,  $T_2$ ,  $T_3$  and  $T_4$ , and are allocated in contiguous processors as follows:  $T_0$  task is assigned in submesh  $\langle 4,0 \rangle \langle 5,2 \rangle$  regardless of the processor in position  $\langle 4,2 \rangle$ , task  $T_1$  is assigned to the sub-mesh  $\langle 2,0 \rangle \langle 3,1 \rangle$ , task  $T_2$  is assigned the submesh in  $\langle 0,5 \rangle \langle 2,6 \rangle$  regardless of the processor in position  $\langle 2,5 \rangle$ , task  $T_3$  is assigned in submesh  $\langle 0,2 \rangle \langle 1,3 \rangle$ , and task  $T_4$  is assigned in submesh  $\langle 6,3 \rangle \langle 7,4 \rangle$  regardless of the processor in position  $\langle 7,4 \rangle$ . This allocation maximizes the adjacency between processors, and produces an external fragmentation of 8 processors. Now if the system



assigns task  $T_5$ , together with task  $T_1$  or  $T_3$ , all of the free processors will be used, and in opposition to the allocation of the 5 tasks, external fragmentation will be minimized. Thus, the contrast of goals 3 and 5 is produced.

Example 3: Exemplification of the contrast between the objectives of the minimization of task, residence time in the queue and the maximization of communication overhead (objectives 1 and 4), is shown when a large number of tasks are sought to be assigned in the processor mesh, and processors to which tasks are to be assigned are not close enough together or contiguous. This is done to avoid producing very high communication costs. As an example, consider allocating the 5 task set:  $T_0, T_1, T_2, T_3$  and  $T_4$ . The number of allocations made to the mesh is minimized, but if the allocator does not consider assignment of disjoint processors by a previous calculation method of communication overhead, tasks will be assigned disjoint in the mesh, causing adjacency to be minimal and communication costs between tasks to be very high.

**B. Functionality of the proposed method**

Proposal: In time  $t$  a 4X4 processor mesh is had, whose status array is shown in Figure 2, where the number 1 represents the occupied processors that were assigned to a task at time  $t-1$ , and the number 0 represents the free processors that have not been assigned to a task or sub-task.

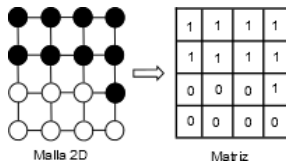


Fig. 2. 4X4 processor mesh represented by a matrix

Symmetrical distances between processors are given in Table I (due to space constraints, only half of the table is shown), these distances represent the “jumps” that a message must execute in order to achieve communication between two processors.

TABLE I. SYMMETRICAL DISTANCES BETWEEN PROCESSORS IN A 4X4 MESH FROM FIGURE 5

	1	2	3	4	5	6	7	8	9
1	0	1	2	3	1	2	3	4	2
2	1	0	1	2	2	1	2	3	3
3	2	1	0	1	3	2	1	2	4
4	3	2	1	0	4	3	2	1	5
5	1	2	3	4	0	1	2	3	1
6	2	1	2	3	1	0	1	2	2
7	3	2	1	2	2	1	0	2	3
8	4	3	2	1	3	2	1	0	4
9	2	3	4	5	1	2	3	4	0

Table 2 shows the waiting queue containing 4 pending execution tasks; said tasks are waiting to be executed in the mesh.

TABLE II. WAITING QUEUE IN TIME T WITH 4 TASKS EACH CONTAINING 4 SUBTASKS

$T_1$	$T_{11}$	$T_{12}$	$T_{13}$	
$T_2$	$T_{21}$	$T_{22}$		
$T_3$	$T_{31}$	$T_{32}$	$T_{33}$	
$T_4$	$T_{41}$			

Previous knowledge is had of the extent of the degree of communication (communication costs), between the main task and sub-tasks that is composed of all tasks that are found in the waiting queue, and the relationship between the same sub-tasks. Table 3, shows the matrix of communication costs for tasks  $T_1$  and  $T_2$ ; Table 4, shows the matrix of communication costs for  $T_3$  and  $T_4$  tasks; communication costs are established between the main task and subtasks and between subtasks. For example, the communication cost between task  $T_1$  and subtask  $T_{11}$  is 3.

TABLE III. MATRIX COMMUNICATION COSTS FOR TASKS  $T_1, T_2$

	$T_1$	$T_{11}$	$T_{12}$	$T_{13}$	$T_2$	$T_{21}$	$T_{22}$
$T_1$	0	3	0	3	0	0	0
$T_{11}$	2	0	1	4	0	0	0
$T_{12}$	0	1	0	2	0	0	0
$T_{13}$	3	5	3	0	0	0	0
$T_2$	0	0	0	0	1	3	0
$T_{21}$	0	0	0	0	1	3	0
$T_{22}$	0	0	0	0	2	0	4

TABLE IV. MATRIX COMMUNICATION COSTS FOR  $T_3, T_4$  TASKS

	$T_3$	$T_{31}$	$T_{32}$	$T_{33}$	$T_4$	$T_{41}$
$T_3$	0	1	3	2	0	0
$T_{31}$	1	0	1	2	0	0
$T_{32}$	4	5	0	1	0	0
$T_{33}$	2	5	2	0	0	0
$T_4$	0	0	0	0	0	3
$T_{41}$	0	0	0	0	2	0

To illustrate the relationship between task and subtasks, consider that you have task  $T_1$  with three sub-tasks  $T_{11}, T_{12}$  and  $T_{13}$ , (as shown in Figure 3). The lines show the transfer of messages, thus task  $T_1$  can send and receive messages from their sub-tasks; in turn sub-tasks can do the same with the main task and each other.

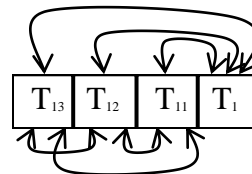


Fig. 3. Message Path between tasks; task with 3 sub-tasks

The functionality of the proposed method is divided into five stages: a) communication between the scheduler and dispatcher, b) dynamic selection of tasks in queue, c) aptitude evaluation of created solutions, d) generation of new populations, and e) allocation of the best individual to the processor mesh. The following sections explain each stage, together with the proposed example.

Communication between the planner and the allocator: Once the allocator counts the number of processors available in the mesh, it reports this amount to the planner; using the example from Figure 2, the allocator will inform the planner of 7 available processors. In our example we have shown a case in which all free processors appear totally adjacent, but if they are found to be disjoint, the process that the method follows is the same.

Dynamic selection of tasks waiting in the queue and the dynamic quadratic assignment of tasks to the processor mesh: With the number of available processors on the grid, the scheduler performs the following three steps to the pre-selection of a set of tasks: 1) dynamically selects a task queue using the ROS. In this method, all the tasks have the same probability of selection [20]. 2) Verifies that number of processors required by the task is less than or equal to the number of processors detected by the allocator; if the condition is met, the number of available processors is reduced by the amount of processors required by the task; if the condition is not met, another task will be randomly selected from the queue. 3) Every time that a task is accepted, three checks are made to the prompt completion of tasks: 1) if whether or not the number of available processors is 0, 2) if the stop condition is true, and 3) if all tasks in the queue have been selected at least once. The Random-Order-of-Service policy, allows the tasks that effectively fit the number of available processors in the net to be selected.

In the example, the first task prompt occurs: a random number is generated, based on the number of jobs in the queue which in this case is 4; if the task can be chosen in the sub-mesh or free submeshes, then it will be considered for calculating its allocation and cost of message transfer; if not, it will choose another task. For this exemplification case, the 2 randomly selected tasks are:  $T_1$  and  $T_2$ ; their placement in the mesh with respective subtasks is shown in Table 5.

TABLE V. LOCATION OF TASKS  $T_1, T_2$  AND  $T_3$  THAT REPRESENT THE FIRST ASSIGNMENT IN THE MESH

1	1	1	1
1	1	1	1
$T_{11}$	$T_{12}$	$T_{21}$	1
$T_1$	$T_{13}$	$T_2$	$T_{22}$

The second assignment, randomly generates a new allocation in the free sub-mesh corresponding to the assignment of tasks  $T_3$  and  $T_4$  (as shown in Table 6).

TABLE VI. TASK ALLOCATION MATRIX ACCORDING TO THE STATE MATRIX OF THE MESH AT TIME T REPRESENTING A SECOND SOLUTION OF THE QUADRATIC DYNAMIC ALLOCATION PROBLEM

1	1	1	1
1	1	1	1
$T_{31}$	$T_{33}$	0	1
$T_3$	$T_{32}$	$T_4$	$T_{41}$

Generation of the third allocation shown in Table 7 produces the assignment of tasks  $T_2$  y  $T_4$  to the mesh.

TABLE VII. MATRIX ASSIGNMENTS ACCORDING TO THE STATE MATRIX OF THE MESH AT TIME T, REPRESENTING A THIRD SOLUTION OF THE QUADRATIC DYNAMIC ALLOCATION PROBLEM

1	1	1	1
1	1	1	1
0	0	$T_{21}$	1
$T_4$	$T_{41}$	$T_2$	$T_{22}$

In the fourth generation, the dynamic task selection produces tasks  $T_2$  and  $T_3$  to the mesh. Produced allocation shown in Table 8.

TABLE VIII. TASK ASSIGNMENT MATRIX ACCORDING TO THE STATE MATRIX OF THE MESH, AT TIME T REPRESENTING A QUARTER SOLUTION OF THE DYNAMIC QUADRATIC ASSIGNMENT PROBLEM

1	1	1	1
1	1	1	1
$T_{31}$	$T_{33}$	$T_{21}$	1
$T_3$	$T_{32}$	$T_2$	$T_{22}$

*Aptitude evaluation of created solutions.* Evaluation of created solutions suitability: At this stage the pre-selected set of tasks in the previous step is evaluated with three different objectives: a) the percentage of external fragmentation (*ef*) produced after allocation in order to minimize the number of idle processors in the mesh. For the example case, the first assignment produces an *ef*=0%, the second allocation 0.14%, the third allocation 0.28% and the fourth allocation 0%. b) The number of tasks that the phenotype assigns to the mesh of processors: In the example, the four allocations manage to position two tasks in the processor mesh. c) Communication Overhead or network contention: The allocation cost is calculated for each task, based on communication costs between tasks and the distances between processors, given the message path from one processor to the other and vice versa.

When considering message passing between processors, one must calculate the cost of transference, from the source to the destination and vice versa. In the exemplified case, the transfer rate from task  $T_1$  to subtask  $T_{11}$  is different than that from sub-task  $T_{11}$  to task  $T_1$ , although both measurements can be equal, the values of the distances between processors remain unchanged.

The values to be calculated are given in the operations shown in Table 9 for task  $T_1$ , in Table 10 for the task  $T_2$ , in Table 11 for the task  $T_3$ , and in Table 12 for the task  $T_4$ . The total of the respective individuals are summed in order to obtain the total solution cost, with a total of 35 for task  $T_1$  (as shown in Table 9) and 17 for task  $T_2$  (shown in Table 10). The representation of the above calculations is given by equation (1).

TABLE IX. CALCULATION OF MESSAGE TRANSFER COST FOR TASK  $T_1$ .

$T_1 \rightarrow T_{11}$	$T_{11} \rightarrow T_1$	$(3+2)*1$	5
$T_1 \rightarrow T_{12}$	$T_{12} \rightarrow T_1$	$(0+0)*2$	0
$T_1 \rightarrow T_{13}$	$T_{13} \rightarrow T_1$	$(3+3)*1$	6
$T_{11} \rightarrow T_{12}$	$T_{12} \rightarrow T_{11}$	$(1+1)*1$	2
$T_{11} \rightarrow T_{13}$	$T_{13} \rightarrow T_{11}$	$(4+5)*2$	18
$T_{12} \rightarrow T_{13}$	$T_{13} \rightarrow T_{12}$	$(2+3)*1$	5
		Total	35

TABLE X. CALCULATION OF MESSAGE TRANSFER COST FOR TASK  $T_2$ .

$T_2 \rightarrow T_{21}$	$T_{21} \rightarrow T_2$	$(1+2)*1$	3
$T_2 \rightarrow T_{22}$	$T_{22} \rightarrow T_2$	$(3+4)*1$	7
$T_{21} \rightarrow T_{21}$	$T_{22} \rightarrow T_{21}$	$(4+3)*1$	7
		Total	17

TABLE XI. CALCULATION OF MESSAGE TRANSFER COST FOR TASK  $T_3$ .

$T_3 \rightarrow T_{31}$	$T_{31} \rightarrow T_3$	$(1+1)*1$	1
$T_3 \rightarrow T_{32}$	$T_{32} \rightarrow T_3$	$(3+4)*1$	7
$T_3 \rightarrow T_{33}$	$T_{33} \rightarrow T_3$	$(2+2)*2$	8
$T_{31} \rightarrow T_{32}$	$T_{32} \rightarrow T_{31}$	$(1+5)*2$	12
$T_{31} \rightarrow T_{33}$	$T_{33} \rightarrow T_{31}$	$(2+5)*1$	7
$T_{32} \rightarrow T_{33}$	$T_{33} \rightarrow T_{32}$	$(1+2)*1$	3
Total			39

TABLE XII. CALCULATION OF MESSAGE TRANSFER COST FOR TASK  $T_4$ .

$T_4 \rightarrow T_{41}$	$T_{41} \rightarrow T_4$	$(3+2)*1$	5
Total			5

The totals obtained from each calculation, add up to make an individual assessment by the value obtained in the objective functions. This step allows the individuals with the best values in each objective function, to be obtained and selected from the population.

Generation of new populations: Once a population has been obtained, the best individuals iteratively build new populations of individuals from which to extract those that best fit, with these, the probabilistic model is estimated.

Estimating the probabilistic model: in this part we will use the simplest probabilistic model, in which all variables describing the problem are independent. We calculate the frequency of task occurrence from a part of the population, containing the best individuals in each empty cell of the mesh at time  $t$ , through truncation selection along with the percentage of the truncation. In this case, the frequency of occurrence can be shown in Table 13, due to space constraints, only the frequencies for processor 0 are shown.

TABLE XIII. FREQUENCY OF OCCURRENCE OF EACH TASK IN EACH CELL.

	P(0,0)	P(0,1)	P(0,2)	P(0,3)	P(1,0)
$T_1$	1	0	0	0	0
$T_{11}$	0	0	0	0	1
$T_{12}$	0	0	0	0	0
$T_{13}$	0	1	0	0	0
$T_2$	0	0	3	0	0
$T_{21}$	0	0	0	0	0
$T_{22}$	0	0	0	3	0
$T_3$	2	0	0	0	0
$T_{31}$	0	0	0	0	2
$T_{32}$	0	2	0	0	0
$T_{33}$	0	0	0	0	0
$T_4$	1	0	1	0	0
$T_{41}$	0	1	0	1	0
$\Sigma$	4	4	4	4	3

Allocation of the best individual to the processor mesh (Determination of the best individual). This step shows the task or tasks that produce the best allocation representing the most feasible solution, and which is assigned to the mesh.

## V. EXPERIMENTS

In this section we explain the experiments conducted with the proposed method, against those of the strict FCFS allocation policy; most of the proposed works use this policy (FCFS) during task planning. At the end of the workload

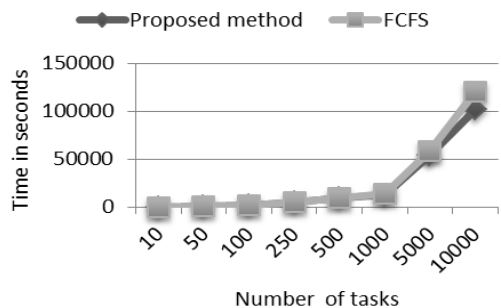
execution in the waiting queue, the five criteria components that are sought to be optimized, in multiprocessor systems are evaluated: utilization, throughput, mean turnaround time, waiting time and total execution time. The parameters that the algorithm uses for its normal operation, that do not need to be optimized are:

- 1) The size of the 2D mesh: This sets the size of the mesh and therefore the number of processors on the target system.
- 2) Number of tasks: the total number of tasks that the system processes also called the overall system load.
- 3) Number of subtasks for each task.
- 4) Time of execution for each task: the parameter that defines the number of seconds, the task will remain within the mesh, constituted by the sum of seconds of each of the subtasks that make up the task.
- 5) Capacity of the queue: the number of tasks that the waiting queue accepts to be processed, and the number of subtasks that each task may contain.
- 6) Number of tasks that the system will seek to enter into the waiting queue: defined as the number of tasks that the algorithm searches for, in the waiting queue using the ROS planning method. The number of tasks is determined by the conditions of the stopping algorithm, whether or not, the tasks waiting in the queue have been selected at least once, or if the number of processors available at time  $t$  was already covered.
- 7) Number of phenotypes or individuals per population that will be created: the parameter that defines the number of individuals, within each one of the populations that constructs the algorithm to determine the best individual (set of tasks assigned to the mesh of processors).
- 8) Number of Populations, that will be created: defined as the number of stocks, that the system generates to extract the best individuals and estimate the probabilistic model.

These are the normal operating parameters of the algorithm. The execution of tests was carried out in the cluster of Liebres InTELigentes servers, consisting of four servers: HP Proliant Quad core with the Linux operating system.

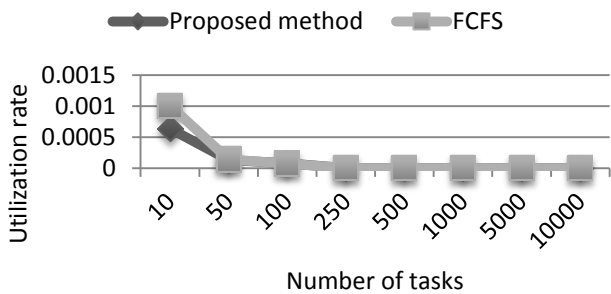
Experiment 1: For the first experiment an 8X8 mesh is used with different queue capabilities: from 10 to 10.000 tasks (as shown in Charts 1 and 5). The number of subtasks per task for this experiment was set 1 to 10 (light load). The task execution time is 1 to 100 seconds. The number of tasks that the system will seek into the waiting queue, once free submeshes are produced in the mesh, is dynamic, and corresponds to the stopping method set in the algorithm as well as the number of phenotypes.

Total execution time: This approach is shown in graph 1. The behavior of both methods, when loads are light, is very similar in this approach. FCFS is a policy free of starvation and non-discriminatory in nature, allowing the task at the head of the queue waiting to be served, once the number of solicited processors is released into the mesh; due to the fact that with light loads, a large number of processors are not required and requests can be quickly met.



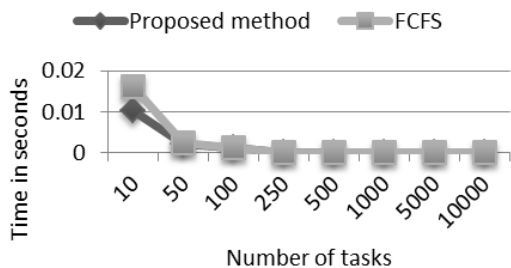
Graph 1. Total execution time

**Utilization:** For this experiment, system utilization is measured by each total workload that the system processes. When comparing system utilization, the behavior of both methods is practically the same as illustrated in graph 2. A previous allocation planning is not synonymous with better system utilization when light loads are processed.



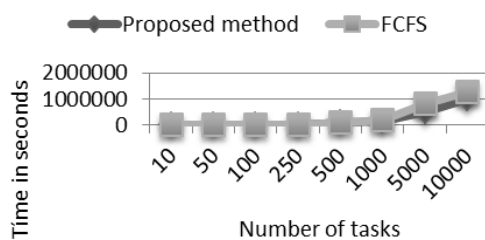
Graph 2. Utilization

**Throughput:** Due to the acceleration that occurs in the allocation, the number of completed tasks per unit of time in the system (when the FIFO allocation policy is used), produces times very similar to the proposed method. That is, the generation of a set of tasks with a small amount of subtasks, upon finalizing execution, enables new tasks to be entered into the mesh without waiting until a large number of processors have been released. The results of this test appear in graph 3.



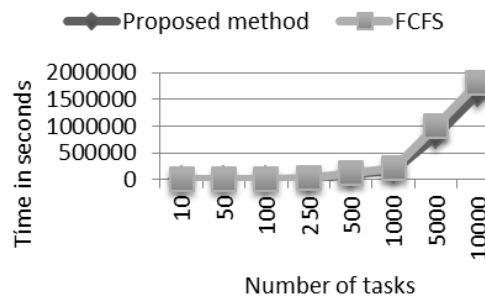
Graph 3. Throughput

**Mean turnaround time:** graph 4 shows how the proposed policy exceeds the FIFO method, when there is an increase of more than 250 tasks in the system load. The significant improvement in time is produced by the response factor, to the task that is at the head of the queue.



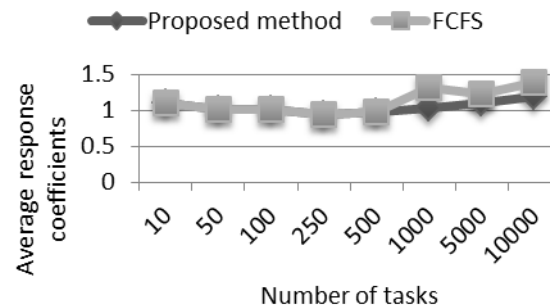
Graph 4. Mean turnaround time

**Waiting time:** The average waiting time of tasks before starting its execution, is significantly improved with the proposed method when the number of tasks in the waiting queue increases (as shown in graph 5). It is assumed that, the improvement in times of this criterion is due, to the utilized ROS planning that does not consider the immediate assignment of the task that is in the head of the waiting queue, but the task search that best suits the free processors.



Graph 5. Waiting time

**Response ratio:** System performance remains constant in both methods when the number of tasks is less than or equal to 500, and varies when loads are increased in the waiting queue (as shown in graph 6). The system performance is considered an important criterion in multiprocessor systems, because it shows the constant and proper use of resources in the target system, or in certain cases, processor waste generated in the target system.

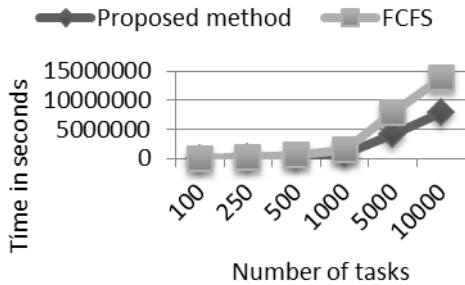


Graph 6. Response ratio

**Experiment 2:** Graphs 7 to 12 show the results of the second experiment. In this experiment, the number of subtasks per task is significantly increased, but the creation of tasks with few subtasks is also allowed. The objective of this experiment is to have a mixture of tasks: tasks with many processor requirements and tasks with little processor requirements. This

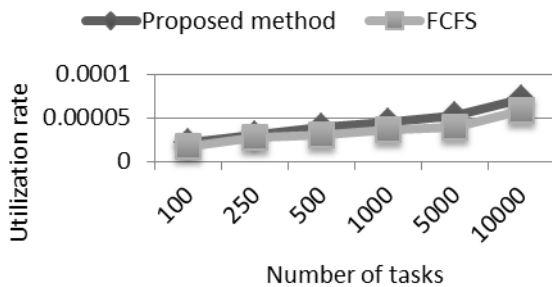
mix allows us to produce external fragmentation of the mesh, and observe the behavior of the algorithm in order to solve the dynamic quadratic assignment problem in the mesh.

**Total execution time:** When the proposed method plans a large quantity of tasks in the waiting queue, you can choose randomly from a variety of task requirements, causing the total execution time to be reduced drastically (as shown in graph 7).



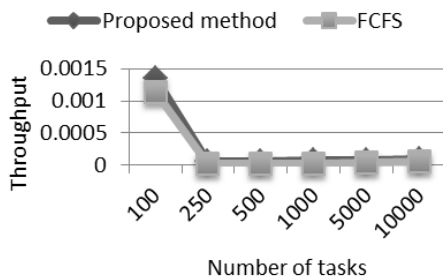
Graph 7. Total execution time

**Utilization:** A better percentage of system utilization is reflected in graph 8, upon using the proposed method due to the fact that in each assignment, all total free processors are assigned to tasks (external fragmentation is decreased). When using a high percentage of processors, network latency increases exponentially because of message passing between tasks.



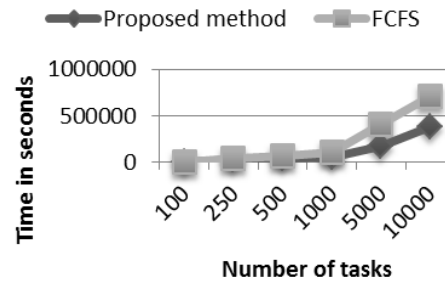
Graph 8. Utilization

**Throughput:** Although the proposed method outperforms the FCFS policy in number of completed tasks per unit time, both provide similar behavior when there is an increase in the number of subtasks per task (as shown in graph 9).



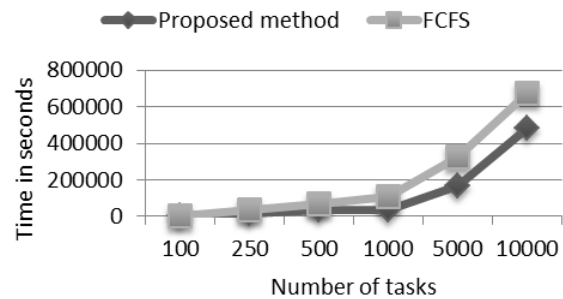
Graph 9. Throughput

**Mean turnaround time:** graph 10 shows the results of mean turnaround time; the proposed method provides shorter times in responses to tasks with a heavy workload.



Graph 10. Mean turnaround time

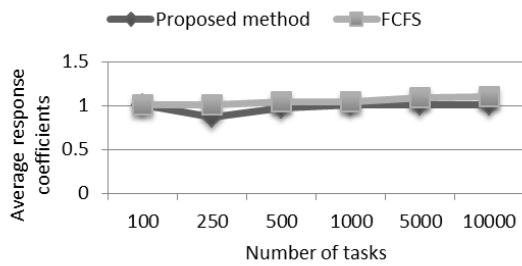
**Waiting time:** The observation in this approach (shown in graph 11), is the reduction in the waiting time of tasks in the waiting queue with the proposed method; the FCFS offers longer waiting times for tasks. Here we observe, the wait that is generated for the tasks with large requirements in the waiting queue; these tasks, must wait until the number of free processors required, to achieve their entry into the mesh are available.



Graph 11. Waiting time

Another important factor is the external fragmentation that occurs when the task at the head of the waiting queue is assigned, and the next task at the head of waiting queue can no longer be allocated due to the number of remaining free processors after the assignment (this being different from the number of processors that is required). Unlike the aforementioned, the proposed method does not assign tasks sequentially in the waiting queue but randomly looks for tasks that best suit the free processors (thereby minimizing task waiting times).

**Response ratio:** As a consequence of the obtained results in the waiting time criterion, system utilization is significantly improved producing better results in the response rate criterion with the proposed method (as shown in graph 12). Achieving a maximum utilization of free processors through planning, every time an assignment is made, yields better results in system utilization.



Graph 12. Response ratio

The goal of both experiments was to show the behavior comparison of the FCFS scheduling policy, which is the most widely, used policy in experiments carried out with the proposed task planning methods in multiprocessor systems. The results show variations in response times of both methods. The main objective of the proposed method is to achieve a pre-planning to the allocation through the optimization of 3 targets, that at the end of workload processing achieves improvement in the established criterion for its evaluation.

## VI. FUTURE WORKS

Considered the basis of this work, future research that arises, is a parallel evaluation of the objectives that are opposed in the planning and allocation of tasks, in a multiprocessor system using multi-core programming. This research is being carried out in a server cluster.

## VII. CONCLUSIONS

This paper, through the joint operation of the task planner and the processor dispatcher considering all the parameters involved in task planning and allocation, presents a strategy that yields a more efficient use of computing resources in a multiprocessor system.

The main objective of this research, is to achieve a pre-planning to the allocation that considers the evaluation of three objective functions that lead to obtaining, a structured assignment avoiding a unique planning of task lists, based on genetic operators or based on the exhaustive search of free submeshes, using geometric models; the proposed method uses the planning policy ROS, whose random behavior allows all tasks to have the same probability of selection, each time the scheduler selects a set of tasks to be assigned to the mesh.

Similarly, the method looks for the best position of the tasks in the processors using a dynamic quadratic assignment, which is determined by estimating and simulating the joint probability distribution, as a mechanism of evolution in order to reduce communication overhead in mesh processors.

The experiments carried out in our work, use the FCFS against the method proposed in this paper. What happens when a planning task with a strict FCFS policy is compared against a totally random policy?

The set of conducted experiments, show the results of the five criteria that are evaluated in multiprocessor systems upon finalizing total execution of workloads, unlike other researches that only seek to optimize a single evaluated criterion. When system loads are light, both planning policies have a similar behavior, and they manage to locate tasks quickly enough in

the mesh, but upon increasing the processor requirements with a larger number of subtasks per task, the proposed method has better results in the following evaluated criteria: utilization, throughput, mean turnaround time, waiting time and the total execution time.

The positivity of the proposed method lies in three key areas: 1) that all tasks have the same probability to be served once a set of tasks are selected for assignment, 2) actively maintain a noncontiguous allocation strategy, which allows it to confront the dynamic quadratic assignment for positioning tasks on processors and 3) avoid producing communication overhead in the processor mesh.

## REFERENCES

- [1] Grama A., Gupta A., Karypis G., and Kumar V. Introduction to Parallel Computing. Second Edition Addison Wesley, January 16, 2003 ISBN: 0-201-64865-2
- [2] Ababneh I., Bani-Mohammad S. "A new window-based job scheduling scheme for 2D mesh Multicomputers". Simulation Modeling Practice and Theory 19 (2011) Pp. 482-493
- [3] Ahmad S. E. "Processors Allocation with Reduced Internal and External Fragmentation in 2D Mesh-based Multicomputers". Journal of Applied Science 11 (6) 943-952, 2011 ISSN 1812-5654 2011 Asian Network for Scientific Information
- [4] Adiga NR., Almasi G., Almasi GS., Aridor Y., et al. "An Overview of the BlueGene/L Supercomputer". Team IBM and Lawrence Livermore National Laboratory
- [5] Bokhari. S. H. "Communication Overhead on the Intel PARAGON, IBM SP2 & MEIKO CS-2". In <http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19960004071.pdf>
- [6] Hussain H. Ur Rehman S., Hamed A., Ullah S., et al. "A survey on resource allocation in high performance distributed computing system". Parallel Computing System & Applications 39 ELSEVIER Pp. 709-736. 2013
- [7] Leung J. Handbook of Scheduling: Algorithms, Models, and Performance Analysis, 1st ed., CRC Press Inc., Boca Raton, FL, USA, 2004
- [8] Walker P., Bundle D., and V. Leung. "Faster high-quality processor allocation". In Proceedings of the 11th LCI International Conference on High-Performance Clustered Computing 2010
- [9] Yoo S., Youn H., and Choo H. "Dynamic Scheduling and Allocation in Two-Dimensional Mesh-Connected Multicomputers for Real-Time Tasks". IEICE TRANS. INF & SYST., VOL. E84-D, No. 5 May 2001
- [10] Chuang P., and Tzeng N. "Allocation precise submesh in mesh connected systems". IEEE Transactions on Parallel and Distributed Systems 5(2) (1994) Pp. 211-217
- [11] Chuang P. and Tzeng N. "An Efficient Submesh Allocation Strategy for Mesh Computer Systems", Proc. Int'l Conf. on Distributed Computing Systems, May 1991, Pp.256-263
- [12] Liu T., Hwang K., Lombardi F., and Bhuyan L. "A Submesh Allocation Scheme for Mesh-Connected Multiprocessor System", Proc. 1995 Int'l Conf. Parallel Processing, vol II, 1995 Pp. 159-163
- [13] Ababneh I. "On submesh allocation for 2D mesh multicomputers using the free-list approach: Global placement schemes". Performance Evaluation 66(2009) ELSEVIER Pp. 105-120
- [14] Velarde A., Ponce de Leon E., and Diaz E. "Planning and Allocation Tasks in a Multicomputer System as a Multi-objective Problem". Advances in Intelligent Systems and Computing 227. EVOLVE 2013. International Conference Held at Leiden University, July 10-13, 2013. Leiden, The Netherlands. Springer
- [15] Feitelson D., Rudolph L., Schwiegelshohn U., Sevcik K., and Wong P. Theory and Practice in Parallel Job Scheduling. Report in <http://www.cs.huji.ac.il/~feit/parsched/jsspp97/p-97-1.pdf>
- [16] Hou E., Ansari N. and Ren H. "A Genetic Algorithm for Multiprocessor Scheduling". IEEE Transactions On Parallel and Distributed Systems, Vol 5, No. 2, February 1994 Pp. 113-120



- [17] Deb K. Multi-objective Optimization Using Evolutionary Algorithm. Wiley 2001.
- [18] Zitzler E., Laumanns M., and Bleuler S. A Tutorial on Evolutionary Multiobjective Optimization In Metaheuristics for Multiobjective Optimization. 2003. In <http://citeseerx.ist.psu.edu/viewdoc/>
- [19] Sinnen O. Tasks Scheduling for Parallel Systems. Wiley Series. 2007
- [20] Rogiest W., Laevens K., Walraevens J., and Bruneel H. "When Random-Order-of-Service outperforms First-Come-First-Served". Operations Research Letters 43. Elsevier. (2015) Pp. 504-506
- [21] Goldberg D. E. "Genetic Algorithms in Search, Optimization and Machine Learning". Addison-Wesley, Reading, MA, 1989
- [22] Kwok Y. K. and Ahmad I. "Efficient Scheduling of Arbitrary Task Graphs to Multiprocessors Using a Parallel Genetic Algorithm". Journal of Parallel and Distributed Computing 47, (1997) Pp. 58-77
- [23] Bohler M., Moore F., and Pan Y. "Improved Multiprocessor Task Scheduling Using Genetic Algorithms" MAICS-99 Proceedings. (1999)
- [24] Hwang R., Gen M., and Katayama H. "A Performance Evaluation of Multiprocessor Scheduling with Genetic Algorithm". Asia Pacific Management Review 11(2) 2006. Pp 67-72.
- [25] Mohamed M. and Awadalla M. "Hybrid Algorithm for Multiprocessor Task Scheduling". International Journal of Computer Science Issues, Vol 8, Issue 3, No. 2, May 2011
- [26] Larrañaga P. "A Review on Estimation of Distribution Algorithms". Estimation of Distribution Algorithms Vol. 2, Kluwer Academic Publishers. 2002. Pp. 57-100
- [27] Das D. and Pradhan D. K. "Job Scheduling in Mesh Multicomputers". IEEE Transactions on Parallel and Distributed Systems, VOL. 9, NO. 1, JANUARY 1998
- [28] Bani-Mohamad S., Ould-Khaoua M., I Ababneh., and Mackenzie L. "Non-contiguous Processor Allocation Strategy for 2D Mesh Connected Multicomputers based on Sub-meshes Available for Allocation". Proceeding of the 12th International Conference on Parallel and Distributed Systems (ICPADS'06) 2006 IEEE
- [29] Bani A. S., "Submesh Allocation in 2D Mesh multicomputers: Partitioning at the Longest Dimension of Request". The International Arab Journal of Information Technology, Vol. 10, No.3, May 2013. Pp. 245 252.
- [30] Y. Zhu. "Efficient processor allocation strategies for mesh-connected parallel computers". Journal of Parallel and Distributed Computing. ELSEVIER. Volume 16, Issue 4, December 1992, Pages 328-337
- [31] G. Kim. "On submesh allocation for mesh multicomputers: a best-fit allocation and a virtual submesh allocation for faulty meshes". Parallel and Distributed Systems, IEEE Transactions on. Volume: 9, Issue: 2. 1998. Pp. 175-185.
- [32] Velarde A."Objective Analysis in Task Planning and Allocation of Multicomputer Systems". RCS Research in Computer Science. Vol 104. ISSN 1870-4069. (2015) Pp. 23-39

# An Improved Image Steganography Method Based on LSB Technique with Random Pixel Selection

Marwa M. Emam

Computer Science Department  
Minia university, Egypt

Abdelmgeid A. Aly

Computer Science Department  
Minia university, Egypt

Fatma A. Omara

Computer Science Department  
Cairo university, Egypt

**Abstract**—with the rapid advance in digital network, information technology, digital libraries, and particularly World Wide Web services, many kinds of information could be retrieved any time. Thus, the security issue has become one of the most significant problems for distributing new information. It is necessary to protect this information while passing over insecure channels. Steganography introduces a strongly approach to hide the secret data in an appropriate media carriers such as images, audio files, text files, and video files. In this paper, a new image steganography method based on spatial domain is proposed. According to the proposed method, the secret message is embedded randomly in the pixel location of the cover image using Pseudo Random Number Generator (PRNG) of each pixel value of the cover image instead of embedding sequentially in the pixels of the cover image. This randomization is expected to increase the security of the system. The proposed method works with two layers (Blue and Green), as (2-1-2) layer, and the byte of the message will be embedded in three pixels only in this form (3-2-3). From the experimental results, it has found that the proposed method achieves a very high Maximum Hiding Capacity (MHC), and higher visual quality as indicated by the Peak Signal-to- Noise Ratio (PSNR).

**Keywords**—Image Steganography; PRNG (Pseudorandom Number Generator); Peak Signal-to-Noise Rate (PSNR); Mean Square Error (MSE)

## I. INTRODUCTION

Data security or data privacy has become increasingly important as more and more systems connected to the internet. In general, protecting the secret messages during transmission becomes an important research issue. To protect secret message during transmission, there are two ways to solve this problem. One way is encryption, which refers to the process of encoding secret information in such a way that only the right person with the right key can decode and recover the original information successfully. Another way is steganography technique, which hides secret information into a cover media or carrier so that it becomes unnoticed, and less attractive [1]. Steganography is one of the important and elegant tools used for securely transfer secret message in an imperceptible manner [2]. The word steganography in Greek means “covered writing” (Greek words “stegos” meaning “cover” and “grafia” means “writing”) [3]. In general, steganography is the art of hiding a message signal into a host signal without any perceptual distortion of the host signal. It hides the secret message within other innocuous looking cover files, called carriers, (i.e. images, text files, audio files, or video files) so

that it cannot be observed [4]. The most frequently used carriers are digital images.

There are two common techniques of embedding in image steganography; *spatial domain* and *transform domain* [5]. According to spatial domain, the secret data or secret message has directly embedded into the LSBs (Least Significant Bit) of image pixels. One of the most known examples of spatial domain method is LSB insertion [6]. In transform domain, the message embedded by modifying frequency coefficients of the cover image [7]. The work in this paper concerns about the spatial domain.

The basic types of steganography are *linguistic* and *technical* steganography. Linguistic steganography takes advantage of the properties of natural language, such as the linguistic structure to hide the messages. While technical steganography is the method of steganography where a tool, device, or method is used to conceal the message and it can be classified into; image, audio, video and text steganography [8].

The basic model of steganography is shown in Fig.1. According to this Figure, steganography process consists of carrier, message, and password. Carrier is also known as cover-object or cover-image, in which message is embedded. The message can be any type of data (plain text, cipher text, or image) that the sender wishes to remain confidential. Password has known as stego-key, which ensures that only recipient who has the stego-key will be able to extract the message from a stego-object. Finally, the cover-object with the secretly embedded message called the stego-object or stego-image [9].

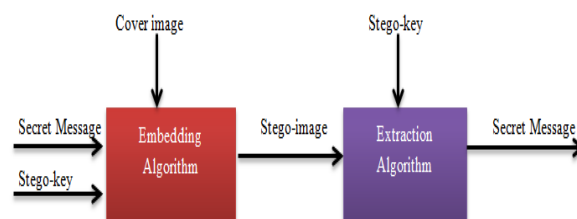


Fig. 1. The Basic model of Steganography

The main goal of stenographic system is the image quality. The Peak Signal-to-Noise Ratio (PSNR) and Mean Squared Error (MSE) metrics are the most common measures used to evaluate the quality of the image [10].

PSNR is one of the metrics to determine the degradation in the embedding image with respect to the cover image. MSE

measures the difference between two images. PSNR and MSE defined in equations 1 and 2 [10].

$$MSE = \left(\frac{1}{MN}\right) \sum_{i=1}^M \sum_{j=1}^N (X_{ij} - X'_{ij})^2 \quad (1)$$

$$PSNR = 10 \log_{10} \frac{I^2}{MSE} \quad (2)$$

Where  $X_{ij}$  is the  $i^{\text{th}}$  row and the  $j^{\text{th}}$  column pixel in the original (cover) image,  $X'_{ij}$  is the  $i^{\text{th}}$  row and the  $j^{\text{th}}$  column pixel in the reconstructed (stego) image,  $M$  and  $N$  are the height and the width of the image,  $I$  is the dynamic range of pixel values, or the maximum value that a pixel can be taken, for 8-bit images;  $I=255$ .

The rest of this paper is organized as follows; related work will discuss in section 2, the proposed method will discussed in details in section 3, performance evaluation will give in section 4. Finally, section 5 concludes the paper.

## II. RELATED WORK

In [11], a modified image steganography method based on LSB technique has presented. This method presents the message by six binary bits by using LSBBraille method (Braille method of reading and writing for blind people) instead of using the ASCII encoding format. This method works with two layers of the RGB image (Blue, and Green layer). The byte of the message is hidden in two pixels only; the first three message bits is hidden in the first pixel and the last three message bits sre hidden in the second pixel by starting with the blue layer then the green layer in the same pixel. According to this embedding way, only two bits for each character of the cover image are changed; 1-bit from blue layer and 1- bit from green layer. In the Blue layer, two bits per pixel are embedded, the message is not only embedded in the first least significant bit (LSB), but also the second least significant bit, and the third least significant bit are allowed to be changed. However, during each process of embedding, only 1-bit of the Blue layer will allowed to be changed. This process is done by taking the last three bits of the Blue layer pixel and applied two equations using the XOR function. Then the third bit of the message in the least significant bit of the Green layer will embedded in the same pixel, and so on as.

In [12], the authors have proposed a secure method of steganography in which three main entities are exercises; pseudo random number generator, Least Significant Bit (LSB) substitution, and Optimal Pixel Adjustment Process (OPAP). In this method, cyclic pixel and indicator technique are used where two channels have been used as data channels and the remaining channel is used as an indicator channel. Red plan has been taken as indicator channel for the first pixel for the subsequent pixels the indicator channels follows a periodic cycle of R, G, and B barring the indicator channel, the other two channels act as the data channels for the corresponding pixels. The pixel intensity determines the bits to be embedded (i.e., LSB's of the indicator channel). If the LSB's of the indicator channel (say R channel) are 00 embed  $k$  bit in G and  $k+1$  bits in B; in case of 01 embed  $k+1$  bits in G and B. if LSB(R) is 10 embed  $k+1$  bits in G and  $k+2$  bits in B. If LSB(R) equals 11 embed  $k+2$  bits in G and  $K+2$  bits in B. A novel 2-key based pseudo random generator is employed

which is used to embed data completely in a unique random fashion based on the user's choice.

Madhu et al [13] have proposed an image steganography method, based on LSB substitution and selection of random pixel of required image area. It generates random numbers and selects the region of the interested where in the required message is embedded along the random pixels. This method is target to improve the security where password is added by LSB of pixels.

In [14], an algorithm called Triple-A algorithm has been proposed. This algorithm uses the same principle of LSB, where the secret message is hidden in the least significant bits of the pixels with more randomization in the selection of the number of used bits and the color channels that are used. This algorithm is divided into two parts (encryption, and hiding). In the hiding part, the RGB image is used as a cover media, which needs to have a pseudorandom number generator. The assumption of PRNG is in every iteration provide two new random numbers as seeds. The seeds of these PRNGs are namely Seed1 (S1), and Seed2 (S2). S1 is restricted to generate numbers in  $[0, 6]$ , while S2 is restricted to the interval  $[1, 3]$ . S1 random number is used to determine the component of the RGB image that is going to be used in hiding the encrypted data. On the other hand, (S2) random number determines the number of the component(s) least significant bits that is used to hide the secret data.

In [15], a spatial domain method has been proposed. The principal of the proposed method is that LSB -3 (Third Least Significant Bit) of the cover image has been used to embed the message bits, and LSB -1, 2 may be modified according to the bits of the message to minimize the difference between the cover and the stego-cover. For more protection to the message bits, a stego-key has been used to permute the message bits before embedding it. However, the results of this method showed that the LSB -1 method has more PSNR values than that the proposed method, which means the LSB -1 image's quality is better than of the modified one, in the same time, the capacity still the same as the modified one.

In [16], a new method is proposed. This method hides the secret message based on searching about the identical bits between the secret message and image pixel value. One pixel of the image is chosen randomly and the image is divided into three layer (Red, Green, and Blue), and then two bits of the secret message are embedded in each layer in the two least significant bit by searching about the identical.

## III. THE PROPOSED METHOD

In this section, the proposed method will present. The proposed method is divided into two algorithms; the embedding algorithm, and the extraction algorithm. The embedding algorithm will plan to hide the byte of the secret message in three pixels only based on randomization in the cover image. It takes the cover image and the secret message characters as an input and converts each byte from the secret message to its binary format using the ASCII encoding format (each byte equal 8-bits). Then, the cover image is converted into three layers (Red, Green, and Blue) layer. Each pixel in the (Blue, and Green) layers is converted to its binary using the

ASCII encoding format. In the embedding technique, (2-1-2) layer is used (i.e. two layers (Blue and Green) are used in the first iteration), in the second iteration only one layer is used (i.e., Blue). In the next iteration, two layers are used (Blue and Green) and so on. The using of two layers then one layer then two layers leads to more secure and getting better PSNR value. The secret message is embedded randomly in the pixel locations using **P**seudo **R**andom **N**umber **G**enerators (**PRNG**) instead of sequential. This method of embedding is considered more secure than the embedding in a sequential manner. The message bits have been embedded in the form of (3-2-3). In which the first 3-bits from the message are embedded in the first random pixel (2-bits in the blue layer at the least and second less significant bit, and 1-bit at the LSB of the green layer). Then, the second 2-bits from the message (fourth and fifth bits) are embedded in the second random pixel at the least and second less significant bit LSB of the Blue layer. After that, the last 3-bits from the message (the sixth, seventh, and eighth bits) are embedded in the third random pixel (2-bits in the blue layer at the least and second less significant bit, 1-bit at the LSB of the green layer), and so on as.

**Pseudo Random Number Generator:**

Pseudorandom number generator acts as a black box, which takes one number (called the seed), and produces a sequence of numbers. The ideal PRNG can generate a unique random integer, to implement the PRNG define a one-to-one function on the integers. Let’s call such function a permutation [17]. It is known in Finite Mathematics that when  $p$  is a prime number,  $x^2 \bmod p$  has some interesting properties. Numbers which are produced by this way are called quadratic residues. The quadratic residues are computed in C using expression (3). In particular, the quadratic residue of  $x$  is unique as long as  $2x < p$ .

For example, when  $p = 11$ , the quadratic residues of 0, 1, 2, 3, 4, 5 are all unique (0, 1, 4, 9, 5, 3) (see Fig. 2).The remaining integers are fitted perfectly into the remaining numbers using expression (4). This only works for prime  $p$ , so new output numbers of 6, 7, 8, 9, and 10 are all unique (8, 6, 2, 7, 10) (see Fig. 3) [17].

$$x * x \% p \quad (3)$$

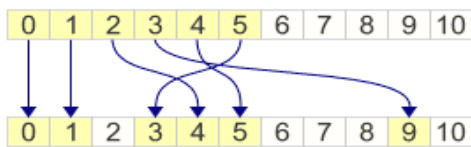


Fig. 2. Quadratic residues of 0, 1, 2, 3, 4, 5 [17]

$$p - x * x \% p \quad (4)$$

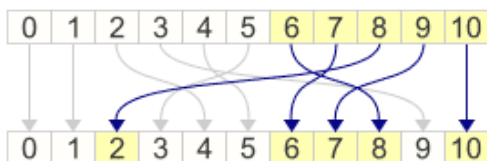


Fig. 3. Quadratic residues of 6, 7, 8, 9, and 10 [17]

*The pseudo code of the proposed method is as follows:*

**Embedding Algorithm:**

**Input:** Cover Image C, Secret Message M, Permutation P

**Output:** Stego Image S.

**Steps:**

1. Split the Cover Image C into three layers Red (R), Green (G), Blue (B).
2. Convert B, and G into blocks;  $B = \{b_1, b_2, b_3 \dots b_n\}$ ,  $G = \{g_1, g_2, g_3 \dots g_n\}$  where each block is only one pixel.
3. Convert each block from B, and G to its ASCII format.
4. Split M into characters,  $M = \{m_1, m_2, m_3 \dots m_n\}$ .
5. Take  $m_i$  from M, and Convert it to its binary format.
6. Declare an array X of integer numbers.
7. For  $i = 1$  to length (X)
8. Generate RN (i) using PRNG generator by applying the equation 3, and 4. Each number of these random numbers will represent the location of the pixels in C.
9. End for
10. Select 8 numbers from RN (i) which located between {1 to 8}, put these numbers in another array called SN.
11. for  $q = 1$  to length (M)\*3
12.  $RS(q) = SN(i) + 8$
13. End for
14. For msgCount=1 to length(M)
15.  $j = 1$
16. Take  $b_i$  from B, and  $g_i$  from G.
17.  $b(RS(j), 7:8) = m_i(\text{msgCount}, 1:2)$
18.  $g(RS(j), \text{end}) = m_i(\text{msgCount}, 3)$
19.  $j++$
20.  $b(RS(j), 7:8) = m_i(\text{msgCount}, 4:5)$
21.  $j++$
22.  $b(RS(j), 7:8) = m_i(\text{msgCount}, 6:7)$
23.  $g(RS(j), \text{end}) = m_i(\text{msgCount}, 8)$
24.  $j++$
25. end for
26. Convert b, and g from binary to decimal.
27. Merge the three layers R, G, B again to construct the stego Image S.

**Example:**

Suppose that the secret message “abc” needs to hide in cover image (Lena 128.bmp). According to the proposed

method, the message characters are converted into its binary format **01100001, 01100010, 01100011**, and then the Blue and Green channels are considered and its pixels are converted into binary format. Each byte from the message will be hidden in three pixels only. Based on the proposed method, it needs to generate nine random numbers to hide the message (message length \* number of pixels). By using the PRNG generator, these random numbers are **RS= {1, 4, 5, 3, 8, 6, 2, 7, 9}**. Therefore, the first byte of the message will be hidden in pixel 1, pixel 4, and pixel 5(i.e., not sequential). Table I shows the embedding process.

TABLE I. RANDOM BASED IMAGE STEGANOGRAPHY EXAMPLE

Pixel no. in sequence manner	Pixels before embedding				Pixels after embedding	
	Blue pixel l	Green pixel	Blue pixel bits	Green pixel bits	Blue pixel bits	Green pixel bits
1	131	137	1000001 1	1000100 1	100000 <u>0</u> <u>1</u>	1000100 <u>1</u>
2	120	135	0111100 0	1000011 1	011110 <u>0</u> <u>1</u>	1000011 <u>1</u>
3	106	132	0110101 0	1000010 0	011010 <u>0</u> <u>1</u>	1000010 <u>1</u>
4	113	131	0111000 1	1000001 1	011100 <u>0</u> <u>0</u>	1000001 1
5	109	129	0110110 1	1000000 1	011011 <u>0</u> <u>0</u>	1000000 <u>1</u>
6	111	131	0110111 1	1000001 1	011011 <u>0</u> <u>1</u>	1000001 <u>0</u>
7	103	127	0110011 1	0111111 1	011001 <u>0</u> <u>0</u>	0111111 1
8	112	133	0111000 0	1000010 1	011100 <u>0</u> <u>0</u>	1000010 1
9	108	122	0110110 0	1000010 0	011011 <u>0</u> <u>1</u>	1000010 <u>1</u>

IV. THE PERFORMANCE EVALUATION

In this section, the proposed method has been tested by taking different messages with different lengths and hiding them in some RGB cover images (i.e., standard images). The proposed method is implemented using MATLAB 11.1.0 software running on a personal computer with a 2.27 GHz Intel (R) Core (TM) I3 CPU , 4 GB RAM and windows 7 as the operating system.

Several experiments with 512 x 512 and 256 x 256 standard images are performed to evaluate the proposed method. Embedding capacity and stego image’s visual quality (PSNR) are used to evaluate the performance of the proposed image steganography method. The results of these experiments are recorded and are summarized in the following tables.

The results in Table II explain that, the proposed method is tested using many images with different capacity, it produces high PSNR and the stego images appears approximately as the cover image that is explained in the MSE. Fig. 4 shows the cover image Lena 512 x 512 and the histograms of its B, and G layers which are used to embed the message given in Table II with capacity (21.845) bytes. Fig. 5 shows the stego image which is obtained after embedding that message and its corresponding histograms.

TABLE II. PSNR AND MSE OF OUR PROPOSED METHOD IN DIFFERENT COVER IMAGES AND DIFFERENT CAPACITY

Cover Image (512 x 512)	Message capacity (bytes)	PSNR of our proposed method (dB)	MSE
Lena	43.690	51.8283	0.4268
	32.768	53.0805	0.3199
	21.845	54.8254	0.2141
Baboon	43.690	51.8436	0.4253
	32.768	53.1012	0.3184
	21.845	54.8664	0.2120
Pepper	43.690	51.8544	0.4243
	32.768	53.1094	0.3178
	21.845	54.8864	0.2111
Bird	65.536	50.0949	0.6362
	43.690	51.8738	0.4224
	32.768	53.1370	0.3158
	21.845	54.8800	0.2114

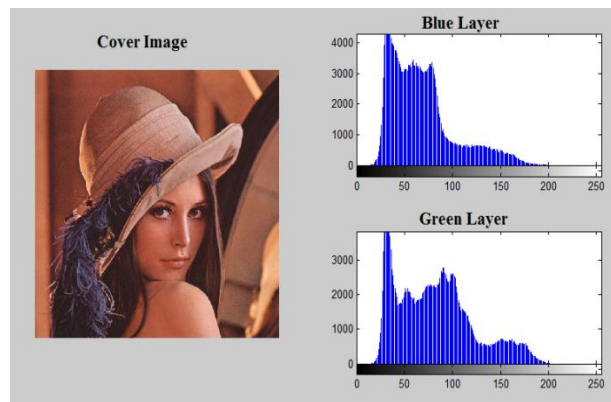


Fig. 4. Cover image Lena and histograms

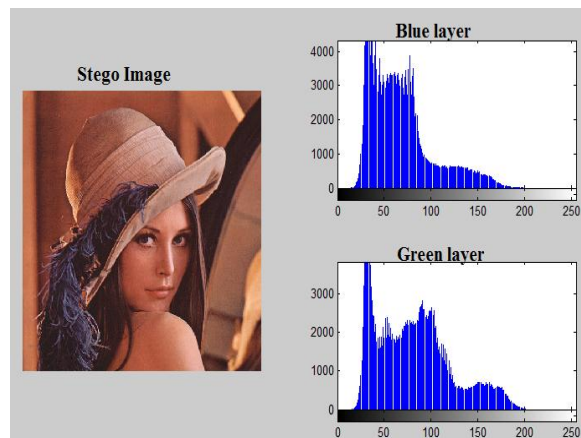


Fig. 5. Stego image Lena and histogram

The results of the comparison study between the proposed method and method in [11] by using different number of characters (bytes) secret message and 512 x 512 cover images (Lena, Baboon, Pepper) are presented in Table III.

TABLE IV. COMPARISON BETWEEN PSNR OF METHOD IN [11] AND OUR PROPOSED METHOD

512 x 512	Cover Images	Message Capacity (bytes)	PSNR (dB)	
			Method in [11]	Proposed method
	Lena	75.836	49.1564	49.4263
	Baboon	82.407	47.7283	49.0871
	Pepper	75.579	47.4422	49.4602

According to the comparison results, it has found that the proposed method has PSNR values better than the method in [11], which mean that the stego image quality of the proposed method will be higher than the stego image quality of the method in [11]. In addition, the average improvement of method in [11] is 48.1089 %, while the average improvement of the proposed method is 49.3245 %. So the proposed method outperforms the method in [11] by 1.22 %. Table IV represents the comparative results of our proposed method and method in [18] using different number of characters bytes (secret message), and 512 x 512 cover images (Lena, Baboon, Airplane).

TABLE V. COMPARISON BETWEEN PSNR OF METHOD IN [18] AND OUR PROPOSED METHOD

512 x 512	Cover Images	Message Capacity (bytes)	PSNR (dB)	
			Method in [18]	Proposed method
	Lena	28.672	43.335	53.6535
	Baboon	28.672	44.902	53.6796
	Airplane	28.672	43.026	53.7138

According to the comparative results, it is found that our proposed method has more PSNR values than that the method in [18], which means that the stego image quality of our proposed method will be higher. In addition, the average improvement of method in [18] is 43.7543 %, while the average improvement of the proposed method is 53.692 %. So the proposed method outperforms the method in [18] by 9.93 %.

Table V represents the comparison of the proposed method and the method in [19] by hiding (145,787 - 144,916 - 145,995) secret bits in 512 x 512 cover images (Lena, Baboon, Pepper) respectively.

TABLE VI. COMPARISON BETWEEN PSNR OF METHOD IN [19] AND OUR PROPOSED METHOD

512 x 512	Cover Images	Message Capacity (bits)	PSNR (dB)	
			Method in [19]	Proposed method
	Lena	145.787	42.26	55.6199
	Baboon	144.916	38.44	55.6750
	Pepper	145.995	42.28	55.6657

According to the results in TABLE V, it is found that the proposed method has more PSNR values than method in [19], which means that the stego image quality of the proposed method is higher than the stego image quality of this method. In addition, the average improvement of method in [19] is 40.993 %, while the average improvement of the proposed method is 55.6535%. So the proposed method outperforms the method in [19] by 14.66 %.

### V. CONCLUSION AND FUTURE WORK

In this paper, a new Steganographic method has proposed, which provides high embedding capacity and PSNR. In addition, by using Pseudo Random Number Generator (PRNG), the security of the system has improved. Experimental results showed that our proposed method is considered an effective Steganographic method while it satisfies the Steganographic system goals.

In the future work, we are looking forward to try applying the proposed method on audio and video. Also, we are looking forward to enhance the proposed method to make the capacity higher than it while keeping the same PSNR or higher.

### REFERENCES

- [1] A. A. Ali and A. H. Seddik, "Image Steganography Technique By Using Braille Method of Blind People (LSBraille)", International Journal of Image Processing (IJIP), Vol. 7, Issue 1, PP. 81-89, 2013.
- [2] A. Ahmed, N. Agrawal, and S. Banerjee, "Image steganography by closest pixel-pair mapping", IEEE- International Conference On Computing, Communications and Informatics (ICACCI), PP. 1971 - 1975, 24-27 Sept. 2014.
- [3] A. Nag, S. Ghosh, S. Biswas, D. Sakar, and P.P. Sakar, "An Image Steganography Technique using X-Box Mapping", IEEE- International Conference On Advances In Engineering, Science and Management(ICAESM-2012), Vol. 3, Issue 12, PP. 709-713, March 2012.
- [4] A. A. Ali and A. H. Seddik, "New Image Steganography Method By Matching Secret Message With Pixels Of Cover Image (SMM)", International Journal of Computer Science Engineering and Information Technology Research (IJCSITR), Vol. 3, Issue 2, PP. 1-10, Jun 2013.
- [5] S. Nazari, A-M. Eftekhari, and M. Sh. Moin, "Secure Information Transmission using Steganography and Morphological Associative Memory", International Journal of Computer Applications, Vol. 61, No. 7, PP. 23-29, January 2013.
- [6] Y. K. Jain and R. R. Ahirwal, "A Novel Image Steganography Method with Adaptive Number of Least Significant Bits Modification Based on Private Stego-Key", International Journal of Computer Science and Security, Vol. 4, Issue 1, PP. 40-94, March 2010.
- [7] S. Sharda and S. Budhiraja, "Image Steganography: A Review", International Journal of Emerging Technology and Advanced Engineering (IJETA), Vol.4, Issue 1, PP. 707-710, January 2013.
- [8] A. A. Ali and A. H. Seddik, "New Image Steganography Method By Matching Secret Message With Pixels Of Cover Image (SMM)", International Journal of Computer Science Engineering and Information Technology Research (IJCSITR), Vol. 3, Issue 2, PP. 1-10, Jun 2013.
- [9] M. E. Saleh, A. A. Ali, and F. A. Omara, "Enhancing Pixel Value Difference (PVD) Image Steganography by Using Mobile Phone Keypad (MPK) Coding", International Journal of Computer Science and Security (IJCSS), Vol. 9, Issue 2, PP. 96-107, 2015.
- [10] A. Almohammad, "Steganography-based Secret and Reliable Communications: Improving Steganographic Capacity and Imperceptibility", Doctor of Philosophy Thesis, Department of Information System and Computing, Brunel University, August 2010.
- [11] M. M. Emam, A. A. Ali, and F. A. Omara, "A Modified Image Steganography Method based on LSB Technique", International Journal of Computer Applications (IJCA), Vol. 125, No. 5, PP. 12-17, 2015.



- [12] R. Amirtharajan, R. subrahmanyam, J. N. Teja, K. M. Reddy, and J. B. B. Rayappan, "Pixel Indicated Triple Layer: A Way for Random Image Steganography ", Research Journal of Information Technology, Vol. 5, Issue 2, PP. 87-99, 2013.
- [13] V. Madhu Viswanatham and J. Manikonda, "A Novel Technique for Embedding Data in Spatial Domain", International Journal on Computer Science and Engineering, IJCSE, vol. 2, PP. 233-236 , (2010).
- [14] A. Gutub, A. A. Qahtani, and A. Tabakh, "Triple-A: Secure RGB Image Steganography Based on Randomization", IEEE, International Conference on Computer System and Application, PP. 400-403, 2009.
- [15] A. I. Abdul-Sada, " Hiding Data Using LSB-3", J.Basrah Researches (Sciences), Vol. 33, No.4, PP. 81-88, DEC 2007.
- [16] J. J. Roque and J. M. Minguet, "SLSB: Improving the Steganographic Algorithm LSB", WOSIS, PP. 57-66. INSTICC Press, 2009.
- [17] Jeff Preshing, (December 24, 2012), "How to Generate a Sequence of Unique Random Integers", [Online]. Available: <http://preshing.com/20121224/how-to-generate-a-sequence-of-unique-random-integers/>.
- [18] U. Lokhande, and A. K. Gulve, "Steganography using Cryptography and Pseudo Random Numbers ", Internatioal Journal of Computer Applications, Vol. 96, No. 19, PP. 40-45, June 2014.
- [19] J. K. Mandal and D. Das, " Colour Image Steganography Based on Pixel Value Differencing in Spatial Domain ", International Journal of Information Sciences and Techniques (IJIST), Vol. 2, No. 4, PP-83-93, July 2012.

# Developing a Feasible and Maintainable Ontology for Automatic Landscape Design

Pintescu Alina

Faculty of Engineering  
Technical University of Cluj-Napoca  
Cluj-Napoca, Romania

Matei Oliviu-Dorin

Faculty of Engineering  
Technical University of Cluj-Napoca  
Cluj-Napoca, Romania

Boancă Iuliana Păunița

Department of Land Measurements and Science  
University of Agricultural Science and Veterinary Medicine  
Cluj-Napoca, Romania

Honoriu Vălean

Automation Department  
Technical University of Cluj-Napoca  
Cluj-Napoca, Romania

**Abstract**—In general, landscape architecture includes analysis, planning, design, administration and management of natural and artificial. An important aspect is the formation of so-called sustainable landscapes that allow maximum use of the environment, natural resources and promote sustainable restoration of ecosystems. For such purposes, a designer needs a complete database with existing and suitable plants, but no designing tool has one. Therefore it is presented the structure and the development of an ontology suitable for storing and managing all information and knowledge about plants. The advantage is that the format of the ontology allows the storage of any plant species (e.g. live or fossil) and automated reasoning. Ontology is a formal conceptualization of a particular knowledge about the world, through the explicit representation of basic concepts, relations, and inference rules about themselves. Therefore the ontology may be used by a design tool for helping the designer and choosing the best options for a sustainable landscape.

**Keywords**—environment; landscapes; ontology; ontology-based simulation; sustainable landscapes

## I. INTRODUCTION

In Romania studies on the landscape made up in the years 1990-1995 aimed especially theoretical approaches and descriptive works, the landscape is analysed through the interaction of natural components. The study of landscape was frequently included in physical-geographical studies. In Europe and the U.S. separation of theoretical and descriptive landscape analysis took place in 1960 when he moved to the quantitative analysis of specific aspects of the landscape. Thus, in 1980 the landscape starts to emerge Science, studying the organization of landscape, natural and human interrelations between components and their temporal dynamics.

There are several application used in landscape design, applications which make use databases and ontology with certain plants. However, these databases or ontology lack many attributes of the plants and are useless for non-specialists. The designers must make use of their experience and background knowledge to use the existing information. Therefore, the aim of the next article is the design and implementation of ontology

about plants used in sustainable landscape design, along with their properties and relationships.

This paper presents the structure and the development of ontology suitable for storing and managing all information and knowledge about plants. The rest of the paper is organized as follows. Section 2 presents the related work in the field of sustainable landscape design. Section 3 discusses the structure and the development of an ontology suitable for storing and managing all information and knowledge about plants. In Section 4, the knowledge modelling environment in this study is presented. Section 5 contains the process to obtain the ontology. Section 6 discusses the results presented in section 5. Finally, conclusion and future work is presented in Section 7.

## II. RELATED WORK

There are different viewpoints regarding sustainable landscape design and types of ontology. In [1] Selman claimed the fact that landscape has become a major issue in spatial policy both as a sector in its own right, important to outdoor amenity and the leisure economy, and, increasingly, as a basis for framing and managing wider socio environmental systems. This trend reflects two broad schools in sustainable landscape development - one focused on the design and protection of scenic assets and the other emphasizing dynamic multifunctional links between ecosystem services and human well-being.

The social dimension of sustainability has been growing in importance as a criterion for evaluating the viability of projects in the construction sector. Reference [2] presents an ontology that can be employed to provide a systematic articulation to the issues that impinge on the social dimension of sustainability appraisals. The development of the social ontology was a consequence of a research project that explored the tools, metrics and models employed in the evaluation of sustainability within the urban environment.

Creating sustainable landscapes is very important especially in the context of environmental conversion of closed mines ([3]) and landslides ([4]); in the context of the accentuated urbanization and of the negative implication on

environment. There are even studies on the fossil flora by Macovei and Givulescu ([5]). Reference [6] tackles algorithms applied in Environmental sciences. Sustainability has been defined by Park ([7]), in part, as the ecological balance that allows us to meet our needs without compromising the ability of future generations to meet theirs.

Sustainable landscape planning is based on a variety of techniques and systems designed to manage environmental problems caused by different factors (increased urbanization, climate change, greenhouse effect). The systems or techniques of sustainable landscape planning is not addressed and not treats only one environmental problem. These systems are designed, most often to alleviate or treat several negative factors that are co-dependent (urbanization - air pollution - noise pollution – greenhouse effect - urban high temperatures - the increase of the impervious surfaces, the reduction of soil permeability – the increase of the volume and speed of the pluvial runoff runaway water - the pollution of the hydrographical network) ([8], [9]).

Even if the conventional landscapes have also an ecological function, generally focuses on the aesthetic function. Stoecklein said that a good landscape architect should have an extensive knowledge of plant, soils, and sites and eye for design, in order to make the right plant selections and put them artistically into their proper locations ([10]). Currently, the sustainable landscapes and the landscape architecture bring forward the ecological function ([11], [12], [13], [14], [15], [16]). Not only selection and artistic arrangement of the plant counts but also the environmental benefits and their functions according to the issues addressed.

Levinthal has considered the problem of landscape design like the tuning of fitness landscapes on which actors adapt ([17]). He examined how alternative organizational designs influence actors' fitness landscapes and, in turn, the behavior that these alternative designs engender. Reducing interdependencies leads to robust designs that result in relatively stable and predictable behaviors. Reference [18] provided an overview of established and emerging forces that influence landscape design. Management, planning, and design as appreciative systems and regenerative processes require the landscape be addressed as system, and the designer be actively engaged in integrative systems thinking. In "Ecological landscape design and planning" the authors contributed to holistic landscape research, creative landscape design and sustainable landscape planning ([19]).

The reference [20] illustrated that environmental quality is one of the factors that has a direct effect on the health and well-being of people in urban areas. They suggested that the urban environment is characterized by altered climate and water relations, damaged soils, man-made substrates, a specialized flora of native and non-native species, and a strong cultural context.

According to [21] sustainable design has a lot to do with society, economy and environment's principles. Social aspects of sustainability, is in need of community participation. Participation in the design process especially in landscape architecture and design is one of the most important factors which are emphasized in recent years and new theories.

Bata delimited the borders of areas, units or entities like universal problem in the spatial and landscape sciences, particularly when addressing natural systems or socio-ecological systems ([22]). The authors of [23] concluded that the best way to keep pace with the latest thinking in environmental planning and landscape design are: the fundamentals of landscape and environmental planning, nature as the all-embracing framework, topography and earth modeling, treatment of wetlands, waterways, and water bodies. The latest researches in the field of the sustainable landscape were focused also on spatial optimization model for landscape planning and renewable energy ([24], [25]).

Reference [26] described the creation of knowledge sharing system for sustainability science through the application of semantic data modeling. An ontology grounded in description logics was developed based on the ISO 15926 data model to describe three types of sustainability science conceptualizations: situational knowledge, analytic methods, and scenario frameworks. Semantic statements were then created using the ontology to describe expert knowledge expressed in research proposals and papers related to sustainability science and in scenarios for achieving sustainable societies.

An important function in the development of urban areas has their ecological assessment and the methods used ([27], [28]). There are various methods for environmental assessment. For the ecological (sustainable) landscape design are available various solutions of numerical modeling and selection of vegetation - but these solutions are not complete. The success of these landscapes is ensured by the high degree of interdisciplinary involved by design and implementation. Initial studies, numerical modeling and planning are important elements for achieving a viable sustainable landscape.

During the development of sustainable landscaping techniques appeared important tools for assessment, modeling, design and implementation. The inconvenience refers to the restriction of the use of these tools ([29]). These restrictions arise because the sustainable landscapes and the related techniques address local and regional conditions and depend on the existing factors in a particular area.

So, the modeling tools, design and selection of vegetation are often developed according to the existing conditions in a particular region of the world or a country, and cannot be applied only under similar conditions to other regions. The vegetation has the one of the most important roles in the proper functioning of the sustainable landscapes. In order to achieve the specific objectives, of a particular sustainable landscaping technique, vegetation should be chosen to optimize, support and comply with its functions. An advanced and complex ontology that include plants with various characteristics enable accurate selections based on the requirements of the techniques that are to be implemented, on local conditions, and on the functions that are intended to be met by the proposed landscape.

### III. STRUCTURE OF THE ONTOLOGY

Ontology is formal conceptualization of a particular knowledge about the world, through the explicit representation

of basic concepts, relations, and inference rules about themselves ([30], [31]). Domain ontology can be used to provide knowledge support in underlying cognitive processes and inter-relations.

Some reasons why someone would develop an ontology are described by [32]: to share common understanding of the structure of information among people or software agents, to enable reuse of domain knowledge, to make domain assumptions explicit, to separate domain knowledge from the operational knowledge, to analyze domain knowledge. Developing ontology is asking to define a set of data and their structure for other programs to use. Problem-solving methods, domain-independent applications, and software agents use ontology and knowledge bases built from ontology as data.

For example, according to [33] the necessity of creating a legal ontology specific to the Romanian juridical context resides in the fact that, although different state institutions hold various databases storing documents, there is a lack of centralization of these data.

The ontology for sustainable landscape design was developed. This ontology can be used as a basis for some applications in a suite of landscape-managing tools: the application could create the possible garden architecture suggestions for the each potential client. With such ontology as the one proposed in this article, software algorithms may be used to extract and match the suitable plants, such as evolutionary ontology ([34]).

Different definition of ontology provides various facilities. The latest development in defining standard languages OWL ontology is produced by World Wide Web Consortium (W3C). Like RDF, OWL makes it possible to describe concepts but in addition, provides many other facilities. It has a richer set of operators (for example and, or, negation). It is based on a different logical model that allows concepts to be simultaneously defined and described. Complex concepts can be defined based on simple concepts. The logical model allows use reasoners to check whether all the definitions and declarations are mutually consistent and can also acknowledges that fit concepts and under the definition. The function of reasoners helps to maintain a proper hierarchy. This hierarchy is useful when working with classes that have more than one parent. Unlike a mere database, such as many existing ones, an ontology is capable of describing the relationships between entities, e.g. what plants match what conditions, what plants need to be planted around what plants etc. Moreover, ontology is a good platform for reasoning and inference, very well described by [35]. This means that the user has all the tools to query the ontology and find out any existing characteristic about the approached domain.

#### IV. KNOWLEDGE MODELLING ENVIRONMENT

Reference [36] presented methodologies for building ontology from scratch. Reference [37] and [38] proposed the solution for ontology based approaches. Typical ontology - enabled tasks include profile matching, gap analysis and the selection of appropriate learning opportunities.

Ontology is a semantically enriched data model that represents a set of concepts within a domain and the

relationships between those concepts ([39]) OWL ontology has similar components to Protégé frame based ontology. However, the terminology used to describe is different from that used by Protégé. OWL ontology is constructed of Individuals, Properties and classes that have correspondent in Protégé: Instance, Slots and Classes. Individuals represent objects in the domain that was interesting. An important difference between Protégé and OWL is that OWL does not use Unified Name Assumption (UNA), means that two distinct name can be the same individual.

For example, “Queen Elizabeth Rose”, “Queen Elizabeth” or “Elizabeth Rose” assigned the same individual. OWL have explicitly stated that individuals are equally distinctive each other. Individuals are also known as instances and may be referred to as “instances of classes”. The properties are binary relations between individuals like individual property of the two links each other. For example, property can link individual isToxic for Bluebells (in Romanian Viorele) and for individual MayLily (in Romanian Lăcrimioare), or property Content can bind Plants for Flowers. Properties can be reversed. Such as the inverse of isToxic is notToxic. The properties can be limited to have a single value. Also they can be transitive or symmetric. Properties are equivalent slots in Protégé. They are also known as logical description of roles and relationships in UML and other object-oriented notations.

#### V. DEVELOPING THE ONTOLOGY

The Artificial-Intelligence literature contains many definitions of ontology, many of these contradict one another ([32]): ontology is a formal explicit description of concepts in a domain of discourse (classes - sometimes called concepts), properties of each concept describing various features and attributes of the concept (slots - sometimes called roles or properties), and restrictions on slots (facets - sometimes called role restrictions).

There are three basic rules for building ontology. These rules help to make decisions in a variety of situations. There is no one correct way to model a domain - there are always alternatives ([32]). The best solution usually depends on the application that builds the ontology and the extent who can expect. Building ontology is an iterative process.

Ontology together with a set of individual instances for classes constitutes a knowledge base. In reality, there is a fine line where the ontology ends and the knowledge base begins. Classes describe concepts in the domain. OWL classes are interpreted as sets which contain individuals. They are those described using formal descriptions (mathematical) which specifies the requirements for class members. For example, class Colors would contain all the colors of our interest field: Blue, Green, Orange, Purple, Red, White, Yellow, AllColours.

Classes can be organized in super classes – sub class hierarchy, which is known as taxonomy. Subclasses has specialized the super classes. For example, grades Contents Plants - Plants can be a subclass of Content class (Content is super class and Plants is class). This can be translated as: “All plants are Content”, “Plants All class members are members of the class Content”. The expression *is a* relation forms the basic structure of the ontology.

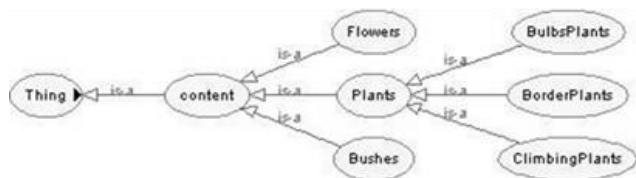


Fig. 1. Relation between classes Content – Plants

If the expression is *A is a B*, it means that node A is a subtype of node B. For example, *BulbsPlants* is a *Plants* class, or *Plants* is a *Content*. It should be noted that *is a* does not mean *is an instance of*. An *instance*, ontologically speaking, is a specific example of something; e.g. a *BulbsPlants* is a *Plants*, but *Gladiola* is an instance of a *BulbsPlants*, rather than a subtype of *BulbsPlants*.

However, if it's know that *BulbsPlants is a Plants*, it is similar with every instance of *BulbsPlants is a Plants*, more over even every instance of *BulbsPlants is a Content*.

Although there is no binding agreement on the appointment of OWL classes, it is recommended that all class names begin with a capital letter and should not contain spaces, for example:

- Garden,
- ThunbergiaAlata,
- BulbsPlants,
- BorderPlants,
- LongBlooming.

Alternatively, it can use the underscore character to group words *Hedera\_iedera*, *Bluebells\_Viorele*.

Empty ontology contains a class called owl: Thing. OWL classes are interpreted as sets of individuals and are constructed from descriptions of the conditions that must be met by an individual to closely match a member of class. OWL ontology is a set of axioms, which provide explicit logical assertions about three types of things: classes, individuals and properties.

Using reasoner it can infer other facts which are inevitably contained in the ontology, for example if an individual *Anemone* is in class *BulbsPlants*, and the class *BulbsPlants* is a subclass of the class *Plants*, a reasoner will infer that *Anemone* is a *Plants*.

It was used the next types of axioms that can be expressed in OWL 2 and also the Manchester syntax:

- *Class declaration* defines a class. A class may contain individuals or other subclass.



Declaration( Class( :Plants ) ) (1)

- *Individual declaration* defines a named individual.



Declaration( NamedIndividual( :Anemone ) ) (2)

- *Class assertion* state that an individual belongs to a class:



ClassAssertion( :BulbsPlants :Anemone ) (3)

- *Subclass assertion* declares that all individuals that belong to a class belong also to another class:



SubClassOf( :BulbsPlants :Plants ) (4)

*Property declaration* defines a data property to link an individual to data, or object property to link to an individual:

Declaration(DataProperty(:SpaceBetweenPlants)) (5)

- *Property assertion* issue the relation of an individual to either data or individual:



DataPropertyAssertion(:SpaceBetweenPlants :Anemone "10"^^xsd:byte) (6)

Before it will be presented the relationships between classes it must have to list the class hierarchy: here was used the Ontograf images produced by Protégé OWL editor. In the plants environment any scene consists of Blooming, Caring, Colors, Content, Shadows, Toxic. These will be the super classes of the Garden ontology (Fig.2).



Fig. 2. Garden ontology super classes

A landscape may refer to Blooming, Caring, Colors, Content, Shadows, Toxic environment. The content has been adapted to these situations. Further classes are customized by introducing appropriate subclasses: Blooming is referred by LongBlooming or ShortBlooming, Caring consists of HardCaring or SimpleCaring, Colors prevailing Blue, Green, Orange, Purple, Red, White, Yellow, AllColours, Content defines Bushes, Flowers and Plants where Plants could be BorderPlants, BulbsPlants and ClimbingPlants species, Shadows might be NaturalLight, Semidarkness and Shadow, Toxic as it may be isToxic, notToxic or SlightlyToxic. The figure below shows a representation of the hierarchy classes in Garden ontology (Fig. 3). It is presents an overall graph of Garden ontology, emphasizing the relationship among the main classes of the ontology. The word concept is sometimes used

instead of classes. Classes are concrete representations of concepts.

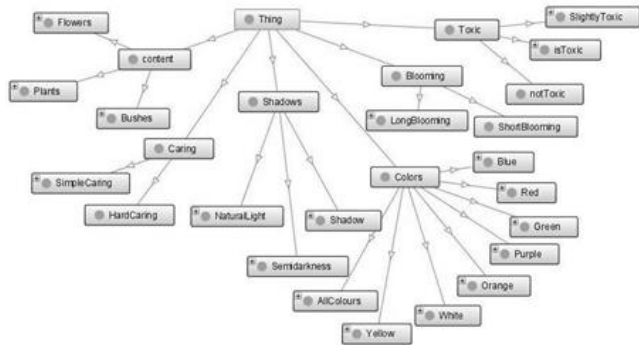


Fig. 3. Class hierarchy

### VI. RESULTS AND DISCUSSION

In this section, it is described the design considerations and modeling concepts, together with a landscape scenario used to illustrate own ideas in that context. The ontology is not complete until the addition of individuals. For every class it can be established as many individuals are necessary. For this ontology it was considered that some individuals for a class are enough. As the individuals are added it was associated the appropriate object and data properties like Figure 4a. In Figure 4b there is an example of individual with description and property assertions. In the Figure 4c was represented all individuals with properties isPerfumed.

For proper experiment it was implemented an ontology with 25 entities (Individuals), 111 relations (Fig. 4a) and 9 data properties. The tests have been carried out using a reasoner and the results have been validated by an expert. On the basis of the description (conditions) a class reasoner can check whether or not that class may have an instance. A class is declared inconsistent if it can have any instance. OWL provides set operations in their usual mathematical meaning.

(a)

(b)

(c)

Fig. 4. (a) Individuals by type; (b) Description and property assertions; (c) Individuals with a property

It is display the set operations on an example in Figure 5. It was defined a class ClimbingPlants which contains four new individuals named SmallBoy, BigBoy, SmallGirl, BigGirl. These individuals must be declared as different from each other, otherwise an OWL reasoner expects that they may be the same:

```

Declaration(Class(:ClimbingPlants))
Declaration(NamedIndividual(:Hedera_Iedera))
Declaration(NamedIndividual(:Ipomoea_Zorele))
Declaration(NamedIndividual(:Lonicera_Caprifoi)))
Declaration(NamedIndividual(:ThunbergiaAlata))
ClassAssertion(:ClimbingPlants:Hedera_Iedera)
ClassAssertion(:ClimbingPlants:Ipomoea_Zorele)
ClassAssertion(:ClimbingPlants:Lonicera_Caprifoi)
ClassAssertion(:ClimbingPlants:ThunbergiaAlata)
DifferentIndividuals(:Hedera_Iedera :Ipomoea_Zorele
:Lonicera_Caprifoi) :ThunbergiaAlata)
    
```

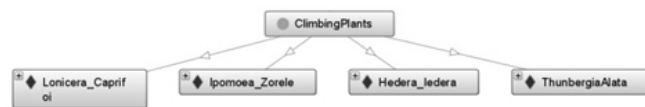


Fig. 5. ClimbingPlants class



For data properties it was necessary to restrict the values of a data property. For example the axiom:

```
ClassAssertion(DataMinCardinality(15:minHeight
xsd:byte):Anemone)
ClassAssertion(DataMinCardinality(3:minSeason
xsd:byte):Anemone)
ClassAssertion(DataMaxCardinality(25:maxHeight
xsd:byte):Anemone)
ClassAssertion(DataMaxCardinality(5:maxSeason
xsd:byte):Anemone)
DataPropertyAssertion(:SpaceBetweenPlants:Anemone
"10"^^xsd:byte)
```

Without having a reasoner is very difficult to maintain large ontology in a state logically correct.

In cases where ontology has classes that have multiple super classes, it is important that every time you build a tree hierarchy of classes that simple. The duty of the reasoner is to calculate and maintain multiple rights (Fig. 6).

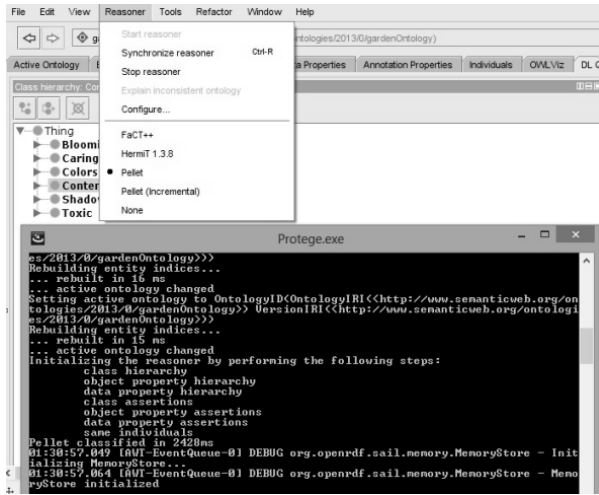


Fig. 6. Garden ontology processed by a reasoner

For validating the ontology, there are testing several sessions of reasoning. The most important constraints ([40]) that can be specified are:

- *allValuesFrom* - which specifies that all values of the properties are in a certain area;
- *someValuesFrom* - which specify that a property has values in a particular field of cardinality;
- *hasValue* - which can be either an individual or a data value. A restriction containing hasValue constraint describes a class of all individuals for which the property concerned has at least one value semantically equal to specified value (it may have other values as well);
- *minCardinality* - describes a class of all individuals that have at least N semantically distinct values (individuals or data values) for the property concerned, where N is the value of the cardinality constraint;

- *maxCardinality* - describes a class of all individuals that have at most N semantically distinct values (individuals or data values) for the property concerned, where N is the value of the cardinality constraint.

In Figure 7 was observed that composing relationships hasDurability and isPerfumed will obtain exactly as result the plants with this properties. It was testing the next query:

$$(\text{hasDurability value true}) \text{ and } (\text{isPerfumed value true}) \quad (7)$$

The result was:

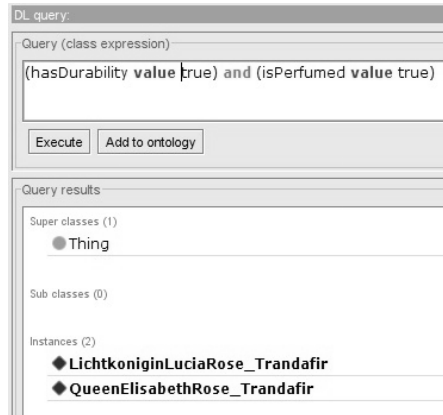


Fig. 7. Perfumed and Durability plants

In order to determine all the bulbs plants which growing after May, the proper query is:

$$\text{BulbsPlants and minSeason min 5} \quad (8)$$

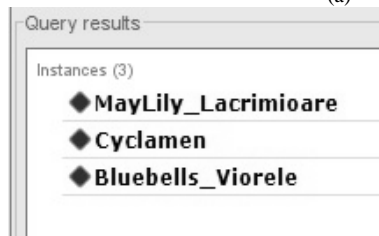
It was obtained as result: Gladiola, Ixia, Dahlia, IrisHollandica, MaryLily (Figure 8a). To find bulbs plants which are Toxic and have property to need space between plants minimum 5 cm the next query was executed:

$$\text{BulbsPlants and isToxic and SpaceBetweenPlants some byte } [>=5] \quad (9)$$

Result is showed in Figure 8b:



(a)



(b)

Fig. 8. (a) Bulbs plants growing after May; (b) Toxic bulbs plants with specified space between plants

To find out the simple caring plants, need natural light and belong one of the next categories: bushes, flowers or climbing plants it is enough to run the next query:

**(Bushes or Flowers or ClimbingPlants) and SimpleCaring and NaturalLight** (10)

The result was:

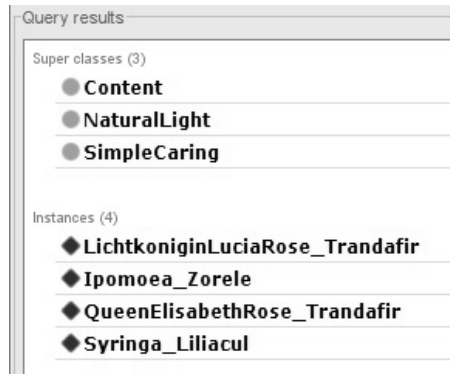


Fig. 9. The simple caring plants, natural light, bushes, flowers or climbing plants

## VII. CONCLUSIONS AND FUTURE WORK

When is necessary to define complex ontology, and beyond, Protégé is recommended. With a highly developed visual interface can define classes, individuals, properties, relations only by a few clicks thus saving much time working with writing effective in OWL.

It was developed genetic algorithms applied to the problems of composition of the sustainable landscapes. Future research can be focused on deepen the study of genetic algorithms and adaptive algorithms to find ways of applying evolutionary computation to solve them. It will mainly consider the transposition of the sustainable landscapes. Theoretical solutions will be integrated into a practical application, usable in real life.

The future research would be interesting to demonstrate that the proposed system is capable of operating in complex areas and can handle large data sets, which can compete with similar systems.

### REFERENCES

[1] P. Selman, "What do we mean by "sustainable landscape"?", *Sustainability: Science, Practice and Policy* 4(2): 23 – 28, 2008.

[2] F.T. Edum-Fotwe and A.D.F. Price, "A social ontology for appraising sustainability of construction projects and developments," *International Journal of Project Management* 27(4): 313-322, 2009.

[3] D. Ciszewski, P. Bijata and K. Klimek, "Reconstruction of post-mining attenuation of heavy metal pollution in sediment of the zlatýpotok, eastern Sudetymts," *Carpathian Journal of Earth and Environmental Sciences*, 9(4): 109 – 120, 2014.

[4] D. Petrea, Ș. Bilașco, S. Roșca, I. Vescan and I. Fodorean, "The determination of the landslide occurrence probability by gis spatial analysis of the land morphometric characteristics (case study: the Transylvanian plateau)," *Carpathian Journal of Earth and Environmental Sciences* 9(2): 91 – 102, 2014.

[5] G. Macovei and R. Givulescu, "The present stage in the knowledge of the Fossil flora at Chiuzbaia," *Maramureș, Romania. Carpathian Journal of Earth and Environmental Sciences* 1(1): 41 – 52, 2006.

[6] S. Hassan, O. Soliman and A. Mahmoud, "Optimized data input for the support vector machine classifier using aster data," *Case study: WadiAtalla area, Eastern Desert, Egypt*, *Carpathian Journal of Earth and Environmental Sciences*, 10(1): 15 – 26, 2015.

[7] S. C. Park, "Sustainable Design and Historic Preservation," *Guiding Principles for Sustainable Design*, NPS Denver Service Center CRM No 2 – 1998, 1993.

[8] C. R. Jacobson, "Identification and quantification of the hydrological impacts of imperviousness in urban catchments: A review," *Journal of Environmental Management* 92: 1438 – 1448, 2011.

[9] T. D. Fletcher and P. H. Andrieu, "Understanding, management and modelling of urban hydrology and its consequences for receiving waters; a state of the art review," *Advances in Water Resources* 51: 261 – 279, 2013.

[10] M. C. Stoecklein, "The Complete Plant Selection Guide for Landscape Design," *Purdue University Press*, West Lafayette, Indiana, ISBN 1-55753-139-0, 2001.

[11] M. Calkins, "Strategy use and challenges of ecological design in landscape architecture," *Landscape and Urban Planning* 73(1): 29 – 48, 2011.

[12] A. Gordon, D. Simondson, M. White, A. Moilanen, and S. A. Bekessy, "Integrating conservation planning and landuseplanning in urban landscapes," *Landscape and Urban Planning* 91(4): 183 – 194, 2009.

[13] M. Scholz, and S. K. Yazdi, "Treatment of road runoff by a combined storm water treatment, detention and infiltration system," *Water Air Soil Pollut.* 198 (1–4): 55 – 64, 2009.

[14] F. Kazemi, S. Beecham, and J. Gibbs, "Streetscale bioretention basins in Melbourne and their effect on local biodiversity," *Ecological Engineering*, 35(10): 1454 – 1465, 2009.

[15] F. Kazemi, S. Beecham, and J. Gibbs, "Streetscape biodiversity and the role of bioretention swales in an Australian urban environment," *Landscape and Urban Planning*, 101(2): 139 – 148, 2011.

[16] S. A. Trowsdale, and R. Simcock, "Urban stormwater treatment using bioretention," *Journal of Hydrology*, 397 (3–4), 3:167 – 174, 2011.

[17] D. A. Levinthal, "Landscape Design: Designing for Local Action in Complex Worlds," *Organization Science*: 342 – 357, 1999.

[18] J. L. Motloch, "Introduction to Landscape Design," ISBN 0-471-35291-8, Printed in the U.S.A, 2000.

[19] J. Makhzoumi, and G. Pungetti, "Ecological landscape design and planning," *Taylor and Francis e-Library*, 2005.

[20] N. Dunnett, and J. Hitchmough, "The dynamic landscape," *British Library Cataloguing, Publication Data*, ISBN 0-203-40287-1 Master e-book ISBN, 2004.

[21] M. Mahdavejajada, and M. Abedi, "Community -oriented landscape design for sustainability in architecture and planning," *Procedia Engineering* 21: 337 – 344, *International Conference on Green Buildings and Sustainable Cities*, 2011.

[22] T. Bata, G. Mezősi, and B. Meyer, "Landscape units for hungary using multiresolution segmentation of geo-data and fuzzy analysis," *Carpathian Journal of Earth and Environmental Sciences*, 9(2): 45 – 56, 2014.

[23] J. O. Simonds, and B. Star, "Landscape Architecture," *Fourth Edition: A Manual of Land Planning and Design*, McGraw Hill book Company (Inc.), New York 6, 2006.

[24] C. W. Huang, Y. P. Lin, T. S. Ding, and J. Anthony, "Developing a Cell-Based Spatial Optimization Model for Land - Use Patterns Planning," *Sustainability* 6, 9139 – 9158, 2014.

[25] R. M. de Waal, and S. Stremke, "Energy Transition: Missed Opportunities and Emerging Challenges for Landscape Planning and Designing," *Sustainability* 6, 4386-4415, 2014.

[26] S. Kraines, and W. Guo, "A System for Ontology-Based Sharing of Expert Knowledge in Sustainability Science," *Data Science Journal* 9: 107 – 123, 2010.

- [27] E. Bakay, "The role of housing estates' green surfaces in forming the city climate of Budapest," *Applied Ecology and Environmental Research* 10(1): 1 – 16, 2012.
- [28] D. Takács, D. K. Varró, and E. Bakay, "Comparison of different space indexing methods for ecological evaluation of urban open spaces," *Applied Ecology and Environmental Research* 12(4): 1027 – 1048, 2014.
- [29] E. Z. Başkent, "Combinatorial Optimization in Forest Ecosystem Management Modeling," *Turkish Journal of Agriculture and Forestry* 25(3): 187-194, 2001.
- [30] J. Biolchini, "Developing a UMLS-based Ontology of Cardiology Procedures for Cognitive Support in Medical Decision Making," *Medical Informatics Training Program, National Library of Medicine*, 2002.
- [31] G. Giray, and M. U. Ünalir, "A method for ontology-based semantic relatedness measurement," *Turkish Journal of Electrical Engineering and Computer Sciences* 21: 420 – 438, 2013.
- [32] N. F. Noy, and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," *Stanford University, Stanford, CA*, 94305, 2001.
- [33] A. Cernian, D. Carstoiu, O. Vasilescu, and A. Olteanu, "OntoLaw - Ontology Based Legal Management and Information Retrieval System," *Journal of Control Engineering and Applied Informatics*, Vol.15, No.4 pp. 77 – 85, 2013.
- [34] O. Matei, D. Contras, and P. C. Pop, "Applying evolutionary computation for evolutionary ontologies," in *Proc. of IEEE Congress on Evolutionary Computation (CEC-2014)*: 1520 - 1527, Beijing, China, 6-11 July, 2014.
- [35] I. Horrocks, and S. Ulrike, "Ontology reasoning in the SHOQ (D) description logic," *International Joint Conference on Artificial Intelligence* 1(3), 2001.
- [36] M. Fernandez-Lopez, and A. Gomez-Perez, "Overview and analysis of methodologies for building ontologies," *Knowledge Engineering Review*, 17(2): 129-156, 2002.
- [37] E. Biesalski, and A. Abecker, "Human Resource Management with Ontologies," *3rd Conference Professional Knowledge Management, Experience and Visions*, Springer, LNCS vol. 3728, 2005.
- [38] M. Laukkanen, and H. Helin, "Competence Manager - Applying the Semantic Web in Practice," *3rd International Semantic Web Conference (ISWC 2004)*.
- [39] S. Furst, R. Blackburn, and B. Rosen, "Virtual team effectiveness: a proposed research agenda," *Information Systems Journal* 9(4): 249 – 269, 1999.
- [40] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, L. A. Stein, and F. W. Olin, "OWL Web Ontology Language Reference," [Online]. Available: <http://www.w3.org/TR/owl-ref/#allValuesFrom-def>, 2004.

# Parallel Implementation of Bias Field Correction Fuzzy C-Means Algorithm for Image Segmentation

Noureddine AITALI  
SSDIA Laboratory, ENSET-Mohammedia  
Hassan II University  
Casablanca, Morocco

Ahmed EL ABBASSI  
PIM Laboratory, FST Errachidia  
Moulay Ismail University, Morocco

Bouchaib CHERRADI  
SSDIA Laboratory, ENSET-Mohammedia  
Hassan II University, Casablanca, Morocco  
Equipe STICE, CRMEF, El Jadida Morocco

Omar BOUATTANE  
SSDIA Laboratory, ENSET-Mohammedia  
Hassan II University  
Casablanca, Morocco

Mohamed YOUSSEFI  
SSDIA Laboratory, ENSET-Mohammedia  
Hassan II University  
Casablanca, Morocco

**Abstract**—Image segmentation in the medical field is one of the most important phases to diseases diagnosis. The bias field estimation algorithm is the most interesting techniques to correct the in-homogeneity intensity artifact on the image. However, the use of such technique requires a powerful processing and quite expensive for big size as medical images. Hence the idea of parallelism becomes increasingly required. Several researchers have followed this path mainly in the bioinformatics field where they have suggested different algorithms implementations. In this paper, a novel Single Instruction Multiple Data (SIMD) architecture for bias field estimation and image segmentation algorithm is proposed. In order to accelerate compute-intensive portions of the sequential implementation, we have implemented this algorithm on three different graphics processing units (GPU) cards named GT740m, GTX760 and GTX580 respectively, using Compute Unified Device Architecture (CUDA) software programming tool. Numerical obtained results for the computation speed up, allowed us to conclude on the suitable GPU architecture for this kind of applications and closest ones.

**Keywords**—Image segmentation; Bias field correction; GPU; Non homogeneity intensity; CUDA; Clustering

## I. INTRODUCTION

In the medical image area, segmentation of anatomical structures is a key step for medical applications such as diagnostics, planning and act operation. Medical images contain a lot of information, and often few of structures are of interest. Segmentation allows visualization of the structures of interest and removing unnecessary information. It also enables structure analysis such as calculating the volume of a tumor, and performing feature-based image-to-patient as well as image-to-image registration, which is an important part of image guided surgery.

In magnetic resonance imaging, the in-homogeneities of intensities, called bias field, are caused by non-uniformities in the Radio Frequency RF field during the acquisition. The result is a shading effect where the pixel or voxel intensities of same tissue class vary slowly over the image domain. This shading can cause severe errors when attempting to segment corrupted images using intensity-based pixel classification methods. It has been shown that this shading is well modeled by the product of the original image and a smooth, very slowly varying multiplier field [1, 2].

In recent literature the authors in [3] proposed modified classical fuzzy c-mean algorithm to be able to get a handle on the intensity in-homogeneities and noisy image effectively. Authors in [4] have demonstrated that the combination of the iterative nonparametric non uniformity normalization and FCM correction method brightens the signal intensity of fatty tissues and that separates the histogram peaks between the fibro glandular and fatty tissues to permit an accurate segmentation between them. In the work of Huan Jun Ding and al [5] they have investigated the feasibility of volumetric breast density quantification with two computer assisted image segmentation methods on medical magnetic resonance imaging (MRI) scans of 40 postmortem breasts.

However, the use of their method is expensive in terms of time execution. In fact, most image segmentation methods proposed in the literature are computationally expensive, especially when run on large medical datasets, and requires powerful hardware to meet desired speed processing.

It also requires several techniques and algorithmic calculation models, which may be sequential or parallel using elementary processors, cellular automata, or neural networks.

In [6] Jaber Juntu and al have discussed two approaches to remedy the problem of the bias field corruption. The one can be used as a preprocessing step where the corrupted MRI image is re-established by dividing it by an estimated bias field signal using a surface fitting approach. The other approach explains how to edit the fuzzy c-means algorithm so that it can be used to segment an MRI image degraded by a bias field signal. The authors in [7] have proposed a fast spatially constrained kernel clustering algorithm in order to segment medical magnetic resonance imaging (MRI) brain images and correct the intensity in-homogeneities. VOVK Uroš and al [8] gave a review of different methods for correction of intensity in-homogeneity in MRI, they classified the methods according to the in-homogeneity correction strategy and different, qualitative and quantitative, evaluation approaches. Earlier, authors in [9] suggested a method that has been applied to 3T and 7T, they have achieved desirable results, and also their method was robust according to initialization. According to all this studies and developed algorithms for image processing, the reach of GPUs has made a revolution in the scientific community by using the power of parallel calculations that are well suited to this type of card.

Graphic processing units (GPUs) were originally created for rendering graphics. Recently, GPUs has emerged as co-processing units for Central Processing Units (CPU) and has become popular for general-purpose high performance computation (GP-GPU) which is mainly attractive and used by many researchers [10-12], to accelerate various digital signal processing applications, including medical image processing [13-15]. GPUs are composed of hundreds of processing cores, highly decoupled, able to achieve immense power of parallel computing. To take advantage of these multi-core architectures, these applications must be parallelized.

Among the images segmentation algorithms with intensity in-homogeneities correction on GPU architecture, authors in [16] proposed an extended mask-based version of the level set method with bias field, recently presented by Li et al. [17]. They develop CUDA implementations for the original full domain and the extended mask-based versions, and compare the methods in terms of speed, efficiency, and performance. The GPU implementation of their version allows a speed up of around 50–100 times for instance, for  $512 \times 512 \times 128$  slices.

Other researchers have suggested different implementation categories of FCM [18] on GPU in order to accelerate the running time. Anderson & al [19] suggested a GPU solution for the Fuzzy C-Means. They have used OpenGL and Cg to achieve approximately two orders of magnitude computational speed-up for some clustering profiles using an NVIDIA8800GPU card. Then they generalized the system for the use of non-Euclidean metrics [20]. On the other and they tried to provide computational intelligence researchers the skills necessary to exploit the low cost and high performance of GPUs with a minimum learning cost [21].

In other works authors implemented a parallel version of FCM on GPU using OpenGL and Cg language [22]. They reached about  $2 \times$  speedup over the sequential implementation.

Rowinska and Goclawski [23] have focused on accelerating the clustering of the FCM algorithm on GPU

using CUDA. They compared their implementation with its C++ sequential implementation, as well as a MATLAB version. In their papers FCM clustering was applied to segment polyurethane foam images. The NVIDIA GeForce GTX560 card was used to perform the parallel experiments, while Intel Core i3 processor was used for the sequential implementation. As a result, they showed that their parallel methods have achieved about  $10 \times$  speed up over the sequential implementation and it was  $50$  to  $100 \times$  faster than the MATLAB version.

In [24] authors proposed a parallel implementation of brFCM which is a faster version of the standard FCM, they tested their algorithm on two GPU cards, Tesla M2070 and Tesla K20m, where the implementation provides about  $2.24 \times$  speedup. The brFCM implementation achieves a speed up of  $23.42 \times$  compared to the traditional FCM. Lately in [25] the authors have introduced a modified FCM algorithm that improved the calculations of the membership matrix and the centers update. Their upgrade version was implemented using CUDA on GPU hardware to rise: the execution time, the visual, and the segmentation efficiency. The experiments used different images of different sizes. They used GTX260 and Intel Core 2 Duo to run the GPU and CPU experiments. The authors achieved at least  $10 \times$  improvements over the sequential FCM version.

Shalom et al. [26] proposed an implementation to improve the computational time of FCM on big data sets using GPU. The practical works were developed on a multidimensional yeast gene expression data set. The researchers stored the distance and membership matrices in the texture memory. The CPU carried out the initial step of the algorithm. Subsequently, the GPU held the most running time parts of the algorithm which are the iterative tasks such as distance calculations, membership calculations, and new cluster centers computations. They developed the experiments on two different GPU cards, GeForce 8500GT and GeForce 8800GTX. The comparison between the parallel and the serial implementations proves an up to  $140 \times$  speedup on 8800GTX according to the CPU implementation. Moreover, the GPU implementation showed an up to  $73 \times$  speedup on 8500GT.

In [27] the researchers implemented and analyzed a parallel dynamic functional connectivity (DFC) algorithm in GPU using two approaches, the first is thread-based and the second is block-based, moreover they have also parallelize the DFC using openMP on fMRI data, they have reached a speedup ranging from  $18.5 \times$  to  $157 \times$  on GPU and  $7.7 \times$  with openMP respectively.

In this paper, our contribution is mainly intended to parallelize the BCFCM algorithm [28] on massively parallel architecture. Our algorithm leads to a better accuracy on segmentation than FCM. But, it is more time consuming. In order to remedy this issue we have exploited the performance of NVIDIA graphic card (GPU) to implement a version that accelerates successfully the BCFCM. We will try later, to detail our method that gives more promising results.

The rest of this paper is organized as follows. In Section II, we summarized the fine-grained parallel model used (GPU of NVidia) and its software development environment (CUDA).

Section III presents a review of the sequential version of the clustering algorithm BCFCM proposed in [28]. Section IV, deals with our parallel implementation for the original and extended algorithms as well as the details of the parallelization. In Section V, we present our findings, compare the results and performance speed-ups. Section VI, concludes the paper and gives some perspectives for this work.

## II. PARALLEL ARCHITECTURE MODEL

In this section we will give a summary presentation of the computational model and the software development environment choosing for the implementation of the algorithm. It is related to GPU massively parallel architecture and it's SDK CUDA.

### A. NVidia GPU Architecture

Modern GPUs are massively parallel processors that support a large number of elementary processors. They are particularly well suited for exploring the calculations on many data types that have high arithmetic intensity. Currently, GPU architecture is composed of an evolving set of processing units flows (SM, SMX Streaming Multiprocessor or: next-generation Streaming Multiprocessor). Such a multiprocessor contains a number of cores (scalar processor), a multithreaded instruction unit, number of registers, local memory and shared memory (Fig.1). The number of processor cores and multiprocessor depends on the architecture and model of the GPU.

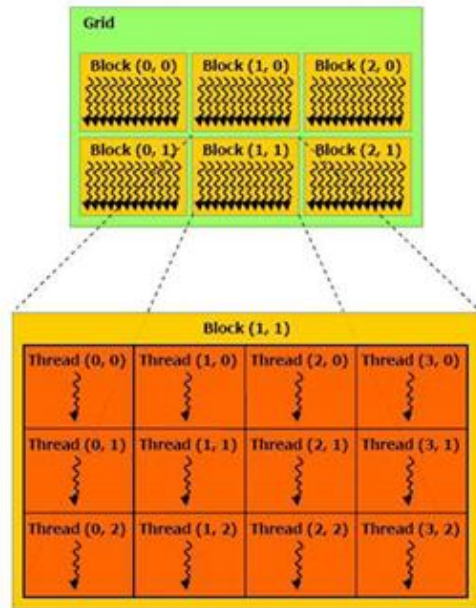


Fig. 1. NVIDIA GPU Grid, Block and Thread hierarchy in CUDA

The fundamental characteristic of the architecture of the GPU is that it has a massively parallel architecture that supports a large number of threads destined to end fine-grained calculation. Each parallelization scheme exploits the ability of mass GPU computing and provides a good load balancing, because each thread executes the same amount of computation.

### B. CUDA: Compute Unified Device Architecture

CUDA is a software development environment based on the C language for GPUs from NVIDIA unveiled in 2007 [29]. It's constituted by a parallel programming model and a set of dedicated instruction. The CUDA codes are compiled using the NVCC compiler [30]. It allows the programmer to define C functions, called kernels, which are executed in parallel by multiple CUDA threads (instantiation of a kernel). The programmer organizes these threads into a hierarchy of grids of thread blocks (Fig.1). A block of threads is a set of concurrent threads that can cooperate with each other through barrier synchronization and shared access. During execution, the threads can access data at different levels of hierarchy: registers, shared memory and global memory. The global memory is accessible by all threads, but its access time is about 500 times slower than the access time to the shared memory and registers.

The treatment of elementary processes (threads) on the GPU is not independent. Indeed, the threads are executed in groups called Warps, where in a warp given (32 threads), all threads execute the same instruction (SIMD).

## III. BACKGROUND: A BIAS FIELD CORRECTION (BC) FUZZY C-MEANS (FCM) ALGORITHM (BCFCM)

The standard FCM [18] objective function for partitioning an MRI image containing  $\{x_k\}_{k=1}^N$  pixels into C clusters is given by:

$$J = \sum_{i=1}^C \sum_{k=1}^N u_{ik}^p \|x_k - v_i\|^2 \quad (1)$$

$u_{i,j}$ : The degree of membership of data  $x_j$  in the cluster  $v_i$ ,

$v_i$ : The prototypes (or center) of the cluster  $i$ ,

$N$ : The total number of pixels in the MRI image

$p$ : A weighting exponent parameter on each fuzzy membership value, it determines the amount of fuzziness of the resulting classification according to :

$$U \left\{ u_{ik} \in [0,1], \sum_{i=1}^C u_{ik} = 1 \quad \forall k, 0 < \sum_{k=1}^N u_{ik} < N, \quad \forall i \right\} \quad (2)$$

The observed MRI signal is modeled as a product of the true signal generated by the underlying anatomy, and a spatially varying factor called the gain field.

$$Y_k = X_k G_k \quad (3)$$

Where  $X_k$  and  $Y_k$  are the true and observed intensities at the  $k^{th}$  pixel, respectively,  $G_k$  is the gain field at the  $k^{th}$  voxel. The application of a logarithmic transformation to the intensities allows the artifact to be modeled as an additive bias field;

$$y_k = x_k + \beta_k \quad (4)$$

Where  $x_k$  and  $y_k$  are the true and observed log-transformed intensities at the  $k^{th}$  pixel, respectively, and  $\beta_k$  is the bias field at the  $k^{th}$  pixel.



Ahmed et al [28] proposed a modification to (1) by introducing a term that allow the labeling of a pixel (Voxel) to be influenced by the labels in its immediate neighborhood. The modified objective function is given by:

$$J_m = \sum_{i=1}^C \sum_{k=1}^N u_{ik}^p \|y_k - \beta_k - v_i\|^2 + \frac{\alpha}{N_r} \sum_{i=1}^C \sum_{k=1}^N u_{ik}^p \left( \sum_{y_r \in N_k} \|y_r - \beta_r - v_i\|^2 \right) \quad (5)$$

Where:

$N_k$ : Set of neighbour's pixels that exist in a window around  $x_k$ .

$N_r$ : Cardinal of  $N_k$

$\alpha$  : Neighbours effect

The new membership function is then given by:

$$u_{ik} = \frac{1}{\sum_{j=1}^C \left( \frac{D_{ik} + \frac{\alpha}{N_r} \gamma_i}{D_{jk} + \frac{\alpha}{N_r} \gamma_j} \right)^{\frac{1}{p-1}}} \quad (6)$$

Where:

$$\gamma_i = \left( \sum_{y_r \in N_k} \|y_r - \beta_r - v_i\|^2 \right) \quad (7)$$

And

$$D_{ik} = \|y_k - \beta_k - v_i\|^2 \quad (8)$$

The cluster prototype (centroid) updating is done by the expression:

$$V_i = \frac{\sum_{k=1}^N u_{ik}^p \left( (y_k - \beta_k) + \frac{\alpha}{N_r} \sum_{y_r \in N_k} (y_r - \beta_r) \right)}{(1 + \alpha) \sum_{k=1}^N u_{ik}^p} \quad (9)$$

The estimated bias field is given by the expression:

$$\beta_k = y_k - \frac{\sum_{i=1}^C u_{ik}^p V_i}{\sum_{i=1}^C u_{ik}^p} \quad (10)$$

Algorithm 1 explains the main steps of this algorithm.

---

**Algorithm 1:** Bias field Correction fuzzy C-Means Algorithm (BCFCM)

---

- 1: Set the parameters  $C$ ,  $p$ ,  $N_r$  and  $\alpha$ .
- 2: Choose the stop criteria: *Error*,
- 3: Initialize the centroids vector  $V$  and estimated bias field  $\beta$ .
- 4: **repeat**
- 5: Update the membership value  $U$  using Eq. (6)
- 6: Update the cluster center vector  $V$  using Eq. (9)
- 7: Update the bias field estimated matrix  $\beta$  using Eq. (10)
- 8: **until**  $\|V_{new} - V_{old}\| < Error$ .

#### IV. PARALLEL FUZZY C-MEANS ALGORITHM FOR BIAS FIELD ESTIMATION AND SEGMENTATION (PBCFCM)

The main objective of the algorithm is to obtain simultaneously a bias field correction and image segmentation. However the expensive computation of this algorithm in its sequential version leads us to exploit the (SIMD) GPU architecture which is the adequate model for this kind of algorithms. Fig.2, illustrates the strategy used to distribute the data, at coarse-grained (Blocks) and fine-grained (threads) levels.

The main idea is to split the image over the GPU so that each pixel presented by one thread can execute its own instructions independently.

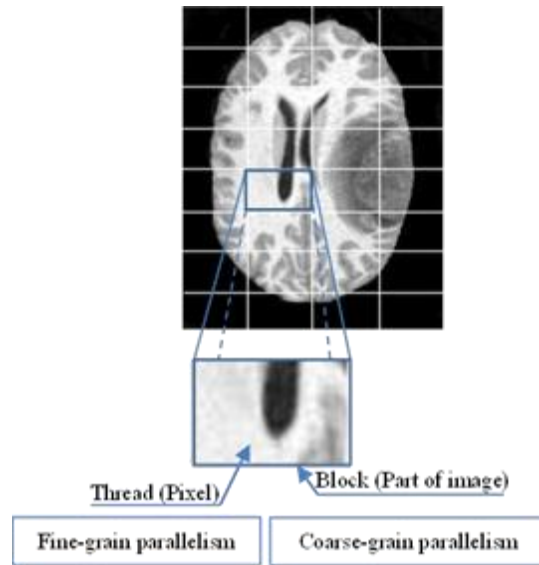


Fig. 2. Data distribution strategy for PBCFCM on GPU

For this algorithm we start with initializing the centroids vector and the variables, then allocate and transfer data from CPU to GPU before the loop iteration. Note that we have used two main kernels, one to compute the membership function and the second to compute the estimated bias field.

Once in the loop, we called the first kernel that computes both operations, the membership function and the expressions for updating centroids. Next step we update the cluster centers vector in CPU, then we transfer the new computed vector to the GPU in order to compute the new estimated bias field, and finally verify the criteria termination based on cluster variation.

The major problem of execution time in our case, is the large data transfer time that eat up the memory. While for small image size, data transfer between CPU and GPU is negligible and can even sometimes be preferable.

In this paper, we present a bias field correction fuzzy c-means implementation which exploits the shared memory for data and constant memory for centroids in addition of registers. While, the part assigned to be executed in CPU is the division operation for the results updating of cluster and

the criteria termination test. Those two operations are assigned to CPU since they do not require processing power. The data transfer cost of cluster centers to constant memory may nearly be ignored compared to the iteration cost. Note that the allocation on device and data transfer are done only one time before starting the loop iteration, to avoid the latency caused by transferring data back and forth from the CPU and GPU.

The Parallel bias field Fuzzy c-means classification algorithm is achieved using the following stages:

---

**Algorithm 2:** Parallel bias field correction Fuzzy C-Means algorithm (PBCFCM).

---

**Stage 1:**

- Initialize the class centers  $V_i$ , ( $i = 1, \dots, C$ ). This is carried out by selecting  $C$  points in the gray level scale  $[0, \dots, 255]$ .
- Read the input data file.
- Initialize the main variables of the algorithm in the CPU.

**Stage 2:** Copy the pixels data, the initial clusters and the estimated bias field to the GPU.

**Stage 3:** Allocate the memory of the membership matrix on the GPU.

**Stage 4:** Compute the membership values based on cluster center for each data member.

**Stage 5:** Compute expression needed for updating clusters on GPU.

**Stage 6:** Update the clusters centers on the CPU.

**Stage 7:** Transfer the new cluster centers to the GPU.

**Stage 8:** Compute the new estimated bias field according to the new clusters on GPU.

**Stage 9:** Check the difference between the current clusters value and its previous one. Test (If the stopping condition is reached) then exit; else return to *Stage 4*.

**Stage 10:** Output the final results.

## V. RESULTS AND DISCUSSION

### A. Experiment setup

Algorithms on host (serial and parallel) were implemented using Microsoft VC++ program and CUDA 7.5 libraries. Sequential BCFCM algorithm is computed to obtain reference runtimes in C and compiled within Microsoft Visual Studio 2013 (Debug mode). Speed-up results were carried out on many devices: Intel(R) Core(TM) i7-4770 8 cores, 3.5GHz (CPU1), Intel(R) Core(TM) i7-4770 8 cores, 2.4 GHz (CPU2), GeForce GT740m, GTX760 and GTX580 GPUs. An example of specifications for tested processor and GPU equipment are summarized in Table 1 and the operating system was 64-bits Windows 7.

### B. Implementation and Validation

To highlight the effectiveness of the proposed method, we extensively experiment both versions, sequential and parallel, of the clustering algorithm on different clinical brain images from online training database BRATS2012 [31] that offer a set of pathological cerebral MRI data. Before this implementation and in order to measure the performance of the proposed implementation in term of time execution, we have used the well-known *Lena* sample image and segment it into its optimal number of clusters which equals to 7 [32] on CPU and

on GPU for different sizes that varies from 1024 to 6553600 pixels. This first assessment stage is done to identify perfectly the behavior of our implantation program on different computational GPU cards. In this section, the parameters set to validate and test the performances of our implementation for *Lena* image are as follow:

$C=7$ ,  $p=2$ ,  $Nr=8$  and  $\alpha =0.85$  ( $\alpha$  represents the neighbors effect as mentioned in [28] for low-SNR images).

To evaluate our implementation, we compared its performance to a sequential equivalent. For this purpose, runtime for serial execution was measured on the host processor that was obtained from single-core execution and referenced (TABLE.I). The GPU based computing duration for the same experiment parameters is compared with single-core timing. In GPU performance evaluation, the allocation on device and data transfer is done only one time before the loop iteration to avoid the latency caused by transferring data back and forth from the CPU and GPU. We calculate the time of the application after the file I/O, in order to show the speedup effect more clearly. The speedup results are normalized to the baseline which is the serial implementation.

TABLE I. HARDWARE SPECIFICATION OF A CPU AND ONE OF THE THREE GPUS (CORE I7, GT 740M) USED IN EXPERIMENTS

Device	Feature	Value
CPU1	Name	i7 4770M
	Frequency	3.5 GHz
	Number of cores	8
	Installed memory	16 GO
GPU1	Name	GT 740
	Multiprocessor count	2
	CUDA cores	384
	Memory bus width	64 bits
	Warp Size	32
	Total global memory	2048MB
	Shared Memory per block	49152 octets
	Max threads per block	1024
	Max threads per multiprocessor	1024
	Memory clock rate	900Mhz

Speed up of the proposed parallel algorithm (PBCFCM) is related to the image size. The larger the size is, the more we get a better speed-up. This rule is not obvious because sometimes a low acceleration can be obtained. Furthermore and according to Fig.3, GPU cards do not have a linear acceleration, we can see that there is almost saturation, but there is always a slight speedup increase as shown in the following Fig.3.

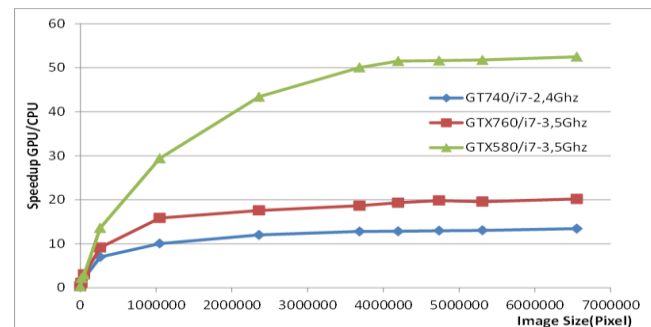


Fig. 3. Evolution of speed-up versus image size on different GPU devices for 32 bit configuration

Experimental results shown above illustrate two behaviors for the three types of GPU cards: GT 740M, GTX 760 and GTX 580, respectively.

Indeed:

- From 1024 pixels to 50526 pixels, the behavior is linear for the three devices,

- From 50526 up to 2359296 pixels the behavior is almost parabolic for GT 740m and GTX 760 while for the GTX 580, the parabolic behavior ends only at the value 3686400 pixels.

Beyond these experimental critical values, we note that, the speed-up tends slowly to the saturation state.

Also note that the maximum speed-up is achieved for the GTX 580 that is almost 5x faster than GT 740M and 2.5x faster than GTX 760.

The speed-ups reached for this figure are 52x, 20x and 13x for GTX 580, GTX 760 and GT 740 respectively for image size up to about to 7 Mega pixels.

According to our study, it seems clear that the performance speed-up depends on the strength of the GPU and the CPU configuration. Fig. 4 confirms the influence of the x64 configuration. Indeed treatment on images to this resolution is better than on x32 configuration (Fig. 3) for the same parallel implementation of the algorithm and different GPU.

Performance improvement is manifested on the three curves of the speed-up to an x64 configuration. The maximum values of speed-up obtained are of the order of 63x, 41x and 19x for GTX 580, GTX 760 and GT 740 respectively for image size up to about to 7 Mega pixels.

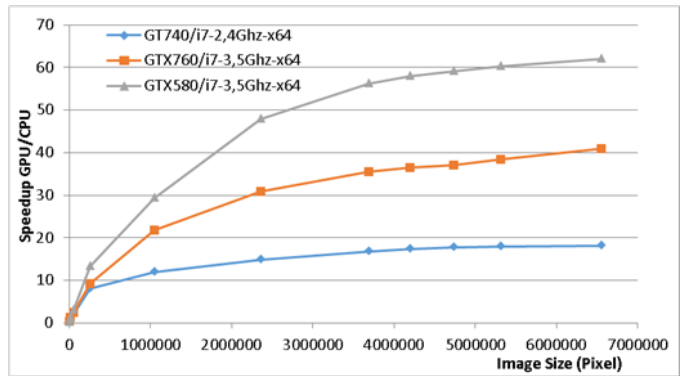


Fig. 4. Evolution of speed-up versus image size on different GPU devices for 64 bit configuration

In most cases and for large image sizes the suitable card having more performance in CUDA language is GTX 580 proved in this study.

To illustrate the impact of the resolution on the same graph, Fig. 5 shows the implementation speed up results of the parallel algorithm PBCFCM over sequential algorithm BCFCM, executed on three GPU cards and confronted the two types of configurations, namely win32 and x64.

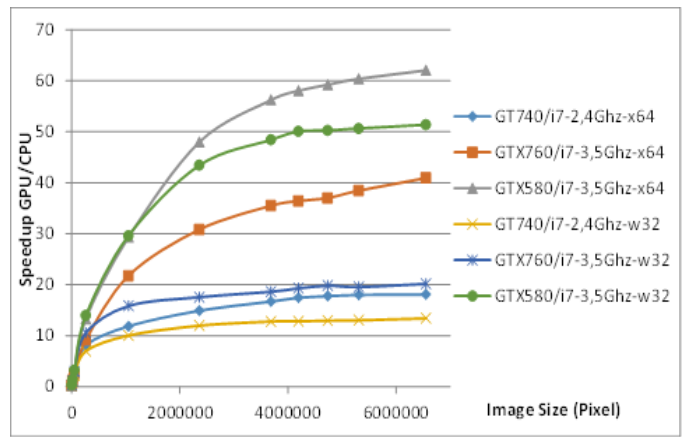


Fig. 5. Speed-up comparison for win32 and win x64 behavior on three GPU devices

Indeed, on the graph of Fig.6, the impact of a very high resolution is mainly shown on the curves where the resolution is x64 for the three types of GPU cards. On the other hand, for the same types of cards and even implementation of algorithm relating to x32 resolution, the speed-up is significantly lower than for x64 resolution. This feature is noticed on the three curves of the graph of figure 6 with x32 resolution. In order to evaluate the performance of the proposed GPU-BCFCM, we present in Fig.6 the variation of the execution time for iteration versus the picture size for the three types of the tested devices.

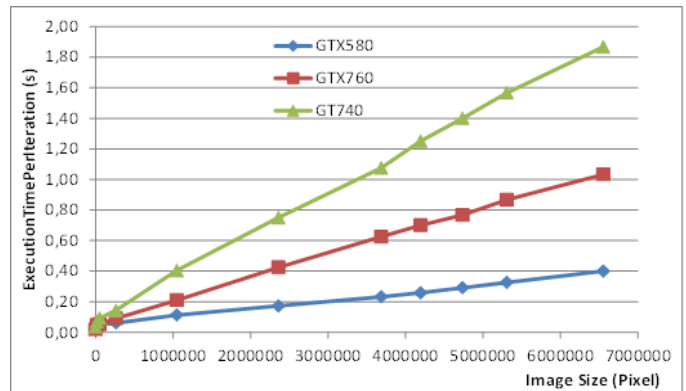


Fig. 6. The variation of execution time per iteration versus image size on different GPU devices

We can clearly see that the variation of execution time per iteration is straight line and strongly related on the performance of the used card. According to our measurements, significant slopes were notified in a decreasing order for: GT740M, GTX 760 and GTX 580 respectively. We found:

$$S_{GT740M} = 2.8 \cdot 10^{-7}$$

$$S_{GTX760} = 1.5 \cdot 10^{-7}$$

$$S_{GTX580} = 6.1 \cdot 10^{-8}$$

All these slopes are defined the ratio:

$$SGTxx = \frac{\text{Execution - Time - Per - Iteration}}{\text{image size on pixels}}$$

We note that the performance test of these cards corresponds to the lower slope. This is given for GTX 580 ( $S_{GTX580}$ ) which has the fastest execution time.

We conclude that GTX 580 still better choose due to its performances compared to GTX 760 or GT 740M related to the slopes of experimental straight lines we introduced in this meaningful study.

The interest of the study that we conducted is clearly manifested for large image sizes. This is validated experimentally for different tests and different GPU cards. This study shows another aspect on performance analysis in terms of data storage effect on the performance of each card (data exchange inside memory card).

Another aspect of GPU performance can be sorted out by inspecting their computation power by testing them by a large amount of computation for a fixed data size. To do so, we extend the measurement for a fixed image size of 1024x1024, and run the FCM algorithm for a number of clusters ranging from 2 to 14 as in Fig.7 that shows the variation of speed-up according to number of clusters for different GPU devices.

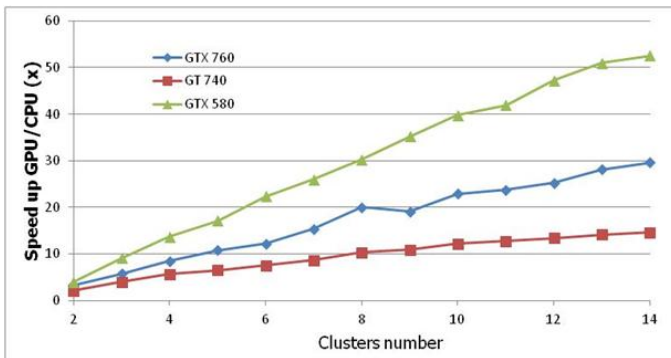


Fig. 7. Speed-up variation versus clusters number for *Lena* image size of 1024x1024

The results investigated in this paper are very relevant as long as they validate the highest importance of GPU processing optimization. This means that how to find optimal material for a given image processing requiring a given amount of memory and powerful ability of computation.

The speed varies almost linearly according to the number of cluster, for GT 740M, GTX 760 and GTX 580, respectively.

In fact, the values minima and maxima are ranging between 2.07x up to 14.64x for GT 740M, and between 3.31x up to 29.63x and finally for GTX 580 the speedup is ranging from 4.01x up to 52.6x.

The curves experimentally defined are close straight lines with dimensionless speed-up are defined per percentage and given as speedup ratios performance (SRP):

$$SRP = \frac{\text{Max speedup for 14 classes}}{\text{Min speedup for 2 classes}} * 100$$

So, for each GPU device we have:

$$SRP_{GT740M} = SC14_{GT740M} / SC2_{GT740M} * 100 = 70.7\%$$

$$SRP_{GTX760} = SC14_{GTX760} / SC2_{GTX760} * 100 = 89.5\%$$

$$SRP_{GTX580} = SC14_{GTX580} / SC2_{GTX580} * 100 = 131.1\%$$

According to these numerical results and parameters performance defined, we conclude that, GTX 580 is faster and powerful than GTX 760 which is also faster than GT740M.

Referring to Fig.8 and Fig.9, the image used for this measurement is 6.5536 million pixels of size. The figure below represents time execution in second versus the used devices categories.

In fact, for serial execution we got as result "62.6 s" which is the highest value comparing to GTX 580 of value 1.2s, GTX760 of value 3.1s and GT740M of value 5.6s respectively (Fig.8).

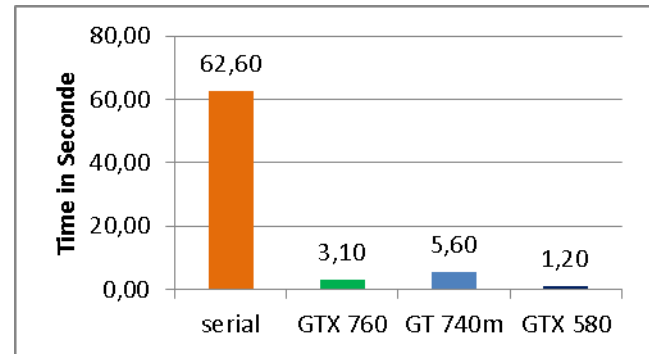


Fig. 8. Comparison of execution time for serial and parallel implementation on three devices

However, after the results expressed experimentally, it is clear that the parallel implementation is very benefic for images ranging from medium-sized to large. On the other hand, Fig.9 shows the speed-up reached in our experimental measurements, for GTX580, GTX760 and GT740M, the values obtained are 52.16x, 20.16x and 11.19x, respectively.

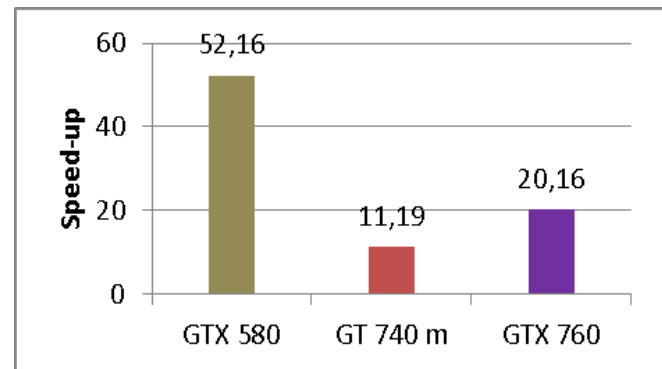


Fig. 9. Speedups of the parallel implementations over the serial one for *Lena* image size of 6.5536 million pixels

### C. Application example

To illustrate the effectiveness of the proposed PBCFCM algorithm we present in fig.10 an application example on T2-weighted MRI image which contains melanoma tumor. This image was segmented into 5 clusters (Background, LCR, GM, WM and Tumor).

Figure 8 (b) shows the estimated Bias field, this was obtained by scaling the bias-field values from one to 255. Figure 8 (c) represents the corrected image after removing the bias field of figure 8(b). Figure 8 (d) shows the segmented image.

Both the parallel and the sequential algorithms was also applied to different MRI datasets image corrupted by bias field artefact and demonstrated its high performance in terms of correction and speed up according to the sequential version.

BCFCM is an algorithm that takes more time to estimate, correct and segment MRI images compared to fuzzy c-mean because it includes additional procedures that make segmentation and correction of intensity in-homogeneity based on neighborhoods. Its variants have high computational complexities that limit their applicability. The large image sizes processing needed by the medical applications motivates the researchers to increase the computation of the enhanced algorithms and take the advantages of the modern GPU of high processing ability.

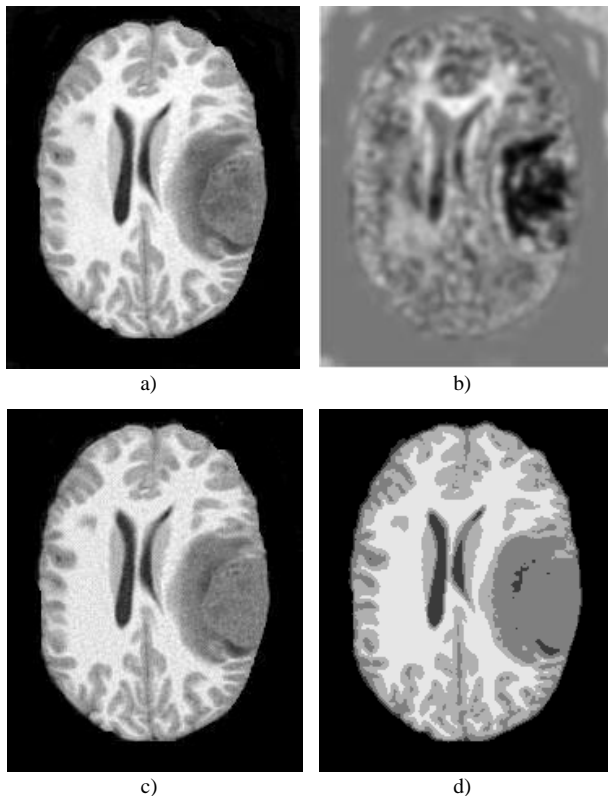


Fig. 10. Example of application. a) Test image, b) Bias-field estimations using PBCFCM algorithm: this was obtained by scaling the bias-field values from one to 255, c) Corrected image, d) 5 clusters segmented image

## VI. CONCLUSION AND PERSPECTIVES

In this paper we presented a parallel implementation of bias field algorithm PBCFCM which is an enhanced version of fuzzy C-means that correct the in-homogeneity intensity and segment the image simultaneously. In fact, many parameters and performance indices are introduced according to the realized measurements on different devices. Our results are translated into meaningful graphs, showing the impact of

architecture on the card and the spot to which it was intended. The speed-up depends on the strength of the GPU, on the image size and on the number of clusters. Larger is the processing amount, the speed-up is higher.

For GTX 580 the speed-up increases until 52x versus the image size and also versus the number of clusters that is limited in this study to  $C=14$ . The obtained curves which are straight lines show that the speed-up is proportional to the number of clusters. For GTX 760 the speed-up increase until 20x versus image size and 29x versus the number of clusters, while for GT740M it reached 11x versus image size and 14x versus the number of clusters.

We can conclude from this study that whatever the performance of the graphic cards, it is necessary to take into account the amount of data to be processed before elaborating any parallel approach. The material performance test procedure remains obvious to fit the physical parallel architecture to the computational model algorithm and data inquiry.

As perspective to this work, we will establish an analytical model for the variation of the speed up versus the data size at first. This model will be improved by introducing the number of clusters especially for MRI classification domain. This study will be specifically launched for the three GPU cards used in our experiments.

## REFERENCES

- [1] Rajapakse, J.C., Giedd, J.N., Rapoport, J.L., 1997. Statistical approach to segmentation of single-channel cerebral MR images. *IEEE Trans. on Med. Imag.* 16, 176-186. doi:10.1109/42.563663
- [2] Unser, M., 1995. Multigrid adaptive image processing. In: *Proceedings of the IEEE Conference on Image Processing*, Vol. I, pp. 49-52. doi:10.1109/ICIP.1995.529036
- [3] KANNAN S R, RAMATHILAGAM S, et PANDIYARAJAN R. Modified bias field fuzzy C-means for effective segmentation of brain MRI. In: *Transactions on computational science VIII*. Springer Berlin Heidelberg, 2010. p. 127-145. doi: 10.1007/978-3-642-16236-7\_9
- [4] LIN, Muqing, CHAN, Siwa, CHEN, Jeon-Hor, et al. A new bias field correction method combining N3 and FCM for improved segmentation of breast density on MRIa). *Medical physics*, 2011, vol. 38, no 1, p. 5-14. <http://dx.doi.org/10.1118/1.3519869>
- [5] DING, Huanjun, JOHNSON, Travis, LIN, Muqing, et al. Breast density quantification using magnetic resonance imaging (MRI) with bias field correction: A postmortem study. *Medical physics*, 2013, vol. 40, no 12, p. 122305. <http://dx.doi.org/10.1118/1.4831967>
- [6] JUNTU, Jaber, SIJBERS, Jan, VAN DYCK, Dirk, et al. Bias field correction for mri images. In : *Computer Recognition Systems*. Springer Berlin Heidelberg, 2005. p. 543-551. doi:10.1007/3-540-32390-2\_64
- [7] LIAO, Liang, LIN, Tusheng, et LI, Bi. MRI brain image segmentation and bias field correction based on fast spatially constrained kernel clustering approach. *Pattern Recognition Letters*, 2008, vol. 29, no 10, p. 1580-1588. doi:10.1016/j.patrec.2008.03.012
- [8] VOVK, Uroš, PERNUŠ, Franjo, et LIKAR, Boštjan. A review of methods for correction of intensity inhomogeneity in MRI. *Medical Imaging, IEEE Transactions on*, 2007, vol. 26, no 3, p. 405-421. doi:10.1109/TMI.2006.891486
- [9] JI, Ze-Xuan, SUN, Quan-Sen, et XIA, De-Shen. A modified possibilistic fuzzy c-means clustering algorithm for bias field estimation and segmentation of brain MR image. *Computerized Medical Imaging and Graphics*, 2011, vol. 35, no 5, p. 383-397. doi:10.1016/j.compmedimag.2010.12.001
- [10] MOHANTY, Saraju P. GPU-CPU multi-core for real-time signal processing. In: *Consumer Electronics*, 2009. ICCE'09. Digest of



- Technical Papers International Conference on. IEEE, 2009. p. 1-2. doi:10.1109/ICCE.2009.5012160
- [11] ZHANG, Kang et KANG, Jin U. Real-time 4D signal processing and visualization using graphics processing unit on a regular nonlinear-k Fourier-domain OCT system. *Optics express*, 2010, vol. 18, no 11, p. 11772-11784. doi: 10.1364/OE.18.011772
- [12] Jararweh Y, JarrahM, Hariri S (2012) Exploiting gpus for compute-intensive medical applications. In:Multimedia computing and systems (ICMCS), 2012 international conference on, IEEE, pp 29–34. doi:10.1109/ICMCS.2012.6320262
- [13] Eklund, A., Dufort, P., Forsberg, D., & LaConte, S. M. (2013). Medical image processing on the GPU—Past, present and future. *Medical image analysis*, 17(8), 1073-1094. doi:10.1016/j.media.2013.05.008
- [14] Pratz, G., & Xing, L. (2011). GPU computing in medical physics: A review. *Medical physics*, 38(5), 2685-2697. <http://dx.doi.org/10.1118/1.3578605>
- [15] SMISTAD, Erik, FALCH, Thomas L., BOZORGI, Mohammadmehdi, et al. Medical image segmentation on GPUs—A comprehensive review. *Medical image analysis*, 2015, vol. 20, no 1, p. 1-18. doi:10.1016/j.media.2014.10.012
- [16] T. Ivanovska, R. Laqua, L. Wang, H. Volzke, and K. Hegenscheid. “Fast Implementations of the Levelset Segmentation Method with Bias Field Correction in MR Images: Full Domain and Mask-Based Versions”. In: *Pattern Recognition and Image Analysis*. Springer Berlin Heidelberg, 2013. p. 674-681. doi:10.1007/978-3-642-38628-2\_80
- [17] Li, C., Huang, R., Ding, Z., et al.: A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI. *IEEE Trans. on Image Processing* 20, 2007–2016 (2011). doi:10.1109/TIP.2011.2146190
- [18] Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2), 191-203. doi:10.1016/0098-3004(84)90020-7
- [19] Anderson, D., Luke, R., Keller, J. (2007). “Speedup of Fuzzy Clustering Through Stream Processing on Graphics Processing Units”, *IEEE Trans. on Fuzzy Systems*. doi:10.1109/TFUZZ.2008.924203
- [20] Anderson, D., Luke, R. H., & Keller, J. M. (2007). Incorporation of non-euclidean distance metrics into fuzzy clustering on graphics processing units. In *Analysis and Design of Intelligent Systems using Soft Computing Techniques* (pp. 128-139). Springer Berlin Heidelberg. doi:10.1007/978-3-540-72432-2\_14
- [21] Anderson, (2008) Parallelisation of Fuzzy Inference on a Graphics Processor Unit Using the Compute Unified Device Architecture, The 2008 UK Workshop on Computational Intelligence, UKCI 2008, pp 1-6.
- [22] Harris, C., & Haines, K. (2005, May). Iterative Solutions using Programmable Graphics Processing Units. In *FUZZ-IEEE* (pp. 12-18).
- [23] Rowińska, Z., & Gocławski, J. (2012). Cuda based fuzzy c-means acceleration for the segmentation of images with fungus grown in foam matrices. *Image Processing & Communications*, 17(4), 191-200. doi: 10.2478/v10248-012-0046-7
- [24] Al-Ayyoub, M., Abu-Dalo, A. M., Jararweh, Y., Jarrah, M., & Al Sa'd, M. (2015). A GPU-based implementations of the fuzzy C-means algorithms for medical image segmentation. *The Journal of Supercomputing*, 1-14. doi:10.1007/s11227-015-1431-y
- [25] Li, H., Yang, Z., & He, H. (2014). An improved image segmentation algorithm based on GPU parallel computing. *Journal of Software*, 9(8), 1985-1990. doi:10.4304/jsw.9.8.1985-1990
- [26] Shalom, S. A., Dash, M., & Tue, M. (2008, November). Graphics hardware based efficient and scalable fuzzy c-means clustering. In *Proceedings of the 7th Australasian Data Mining Conference-Volume 87* (pp. 179-186). Australian Computer Society, Inc.
- [27] AKGÜN, Devrim, SAKOĞLU, Ünal, ESQUIVEL, Johnny, et al. GPU accelerated dynamic functional connectivity analysis for functional MRI data. *Computerized Medical Imaging and Graphics*, 2015, vol. 43, p. 53-63. doi:10.1016/j.compmedimag.2015.02.009
- [28] M.N. Ahmed, N.A. Mohamed, A.A. Farag, T. Moriarty, A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data, *IEEE Trans. Med. Imaging* 21 (2002) 193–199. doi: 10.1109/42.996338
- [29] [http://www.nvidia.com/object/cuda\\_home\\_new.html](http://www.nvidia.com/object/cuda_home_new.html).
- [30] Nvcc: Cuda toolkit documentation. <http://docs.nvidia.com/cuda/cuda-compiler-driver-nvcc/>.
- [31] <http://www2.imm.dtu.dk/projects/BRATS2012/>.
- [32] Ouadfel, S., Batouche, M., & Ahmed-Taleb, A. (2012). ACPSO: A Novel Swarm Automatic Clustering Algorithm Based Image Segmentation. In S. Ali, N. Abbadieni, & M. Batouche (Eds.) *Multidisciplinary Computational Intelligence Techniques: Applications in Business, Engineering, and Medicine* (pp. 226-238). Hershey, PA: Information Science Reference. 5. Chapter 14.

# An Algerian dialect: Study and Resources

Salima Harrat\*, Karima Meftouh†, Mourad Abbas‡, Khaled-Walid Hidouci§ and Kamel Smaili¶

\*Ecole Supérieure d'Informatique (ESI), Algiers, Algeria

†Badji Mokhtar University, Annaba, Algeria

‡CRSTDLA Centre de Recherche Scientifique et Technique  
pour le Développement de la Langue Arabe, Algiers, Algeria

§Ecole Supérieure d'Informatique (ESI), Algiers, Algeria

¶Campus Scientifique LORIA, Nancy, France

**Abstract**—Arabic is the official language overall Arab countries, it is used for official speech, news-papers, public administration and school. In Parallel, for everyday communication, non-official talks, songs and movies, Arab people use their dialects which are inspired from Standard Arabic and differ from one Arabic country to another. These linguistic phenomenon is called diglossia, a situation in which two distinct varieties of a language are spoken within the same speech community. It is observed Throughout all Arab countries, standard Arabic widely written but not used in everyday conversation, dialect widely spoken in everyday life but almost never written. Thus, in NLP area, a lot of works have been dedicated for written Arabic. In contrast, Arabic dialects at a near time were not studied enough. Interest for them is recent. First work for these dialects began in the last decade for middle-east ones. Dialects of the Maghreb are just beginning to be studied. Compared to written Arabic, dialects are under-resourced languages which suffer from lack of NLP resources despite their large use. We deal in this paper with Arabic Algerian dialect a non-resourced language for which no known resource is available to date. We present a first linguistic study introducing its most important features and we describe the resources that we created from scratch for this dialect.

**Keywords**—Arabic dialect, Algerian dialect, Modern Standard Arabic, Grapheme to Phoneme Conversion, Morphological Analysis

## I. INTRODUCTION

Under-resourced languages are languages which lacks resources dedicated for natural language processing. In fact, these languages suffer from unavailability of basic tools like corpora, mono or multilingual dictionaries, morphological and syntactic analyzers, etc. This lack of resources makes working with these languages a great challenge, especially when we deal with unwritten languages like Arabic dialects. Compared to other under-resourced languages, Arabic dialects present the following additional difficulties:

- Since they are spoken languages they are not written and there are no established rules to write them. A same word could have many orthographic forms which are all acceptable since there is no writing rules as reference.
- The flexibility in the grammatical and lexical levels despite their belonging to Arabic Language.
- Besides the fact that these dialects are different from Arabic, they are also different from each other. For instance, dialects of the Maghreb differ from those of

the middle-east. They may be also different inside the same country.

- These dialects are also widely influenced by other languages such as French, English, Spanish, Turkish and Berber.

In Algeria, as well as in all arab countries, these dialects are used in everyday conversations. However, with the advent of the internet they are increasingly used in social networks and forums. They emerge on the web as a real communication language due to the ease to communicate in dialect especially for people with low level of education. But unfortunately basic NLP tools for these dialects are not available.

This work is a first part of the Project TORJMAN<sup>1</sup> which is a Speech-To-Speech Translator between Algerian Arabic dialects and MSA. Unlike Middle-East Arabic dialects, Algerian Arabic dialects are non-resourced languages, they lack all kinds of NLP resources. Consequently, TORJMAN begins from Scratch.

In this paper, we describe and extend resources creation tasks for Arabic dialect of Algeria that appeared in [1] and [2]. We focus on Algiers dialect which is the spoken Arabic of Algiers (capital city of Algeria) and its periphery. This choice is justified by the fact that this dialect is the one we know best and practice since we are native speakers of this dialect. For convenience of reference, we will design Algiers dialect by ALG, this will make this manuscript easier to read.

This paper is organized as follows: before dealing with Algerian dialect we give in Section II a brief overview of Arabic language, whereas in Section III we present different aspects of ALG. The following Sections will be dedicated to the resources that we created, we detail how we made the first corpus of Algiers dialect (Section IV). Then we present ALG grapheme-phoneme converter(Section V) which has allowed us to get a phonetized corpus of Algiers dialect. In Section VI we describe how we created a morphological analyzer for ALG by adapting BAMA[3] the well known analyser for MSA. Finally, we will conclude by summarizing the main ideas of this work and by giving our future tendencies.

## II. ARABIC LANGUAGE

Arabic is a Semitic language, it is used by around 420 million people. It is the official language of about 22 countries. Arabic is a generic term covering 3 separate groups:

<sup>1</sup>TORJMAN is a national research project which is totally financed by the Algerian research ministry, this appellation means translator or interpreter in English.



- Classical Arabic: is principally defined as the Arabic used in the Qur'an and in the earliest literature from the Arabian peninsula, but also forms the core of much literature until the present day.
- Modern Standard Arabic: Generally referred as MSA (Alfus'ha in Arabic), is the variety of Arabic which was retained as the official language in all Arab countries, and as a common language. It is essentially a modern variant of classical Arabic. Standard Arabic is not acquired as a mother tongue, but rather it is learned as a second language at school and through exposure to formal broadcast programs (such as the daily news), religious practice, and newspaper [4].
- Arabic dialects: also called colloquial Arabic or vernaculars are spoken varieties of Arabic language. In contrast to classical Arabic and MSA, they are not written. These dialects have mixed form with many variations. They are influenced both by the ancient local tongues and by European languages such as French, Spanish, English, and Italian.<sup>2</sup> Differences between these variants of spoken Arabic throughout the Arab world can be large enough to make them incomprehensible to one another. Hence, regarding the large differences between dialects, we can consider them as disparate languages depending on the geographical place in which they are practiced. Thus, most of the literature describe Arabic dialects from the viewpoint of east-west dichotomy [5]:<sup>3</sup>
  - Middle-east dialects: include spoken Arabic of Arabian peninsula(Gulf countries and Yemen), Levantine dialect (Syria, Lebanese, Palestinian and Jordan), Iraqi dialect Egyptian and Sudan dialect.
  - Maghreb dialects: Spoken mostly in Algeria, Tunisia, Morocco, Libya and Mauritania. Note that, Maltese a form of Arabic dialect is most often found in Malta.

In the next section, we will focus on a Maghreb dialect from Algeria and more specifically the dialect spoken in Algiers the capital city of Algeria, we will highlight its most features in contrast to MSA.

### III. SPECIFICITIES OF ALGIERS DIALECT

Algiers dialect (ALG) is the dialectal Arabic spoken in Algiers and its periphery. This dialect is different from the dialects spoken in the other places of Algeria. It is not used in schools, television or newspapers, which usually use standard Arabic or French, but is more likely, heard in songs if not just heard in Algerian homes and on the street. Algerian Arabic is spoken daily by the vast majority of Algerians [7]. ALG as the other Arabic dialects simplifies the morphological and syntactic rules of the written Arabic. In [8], the author draws how match spoken Arabic is different from written

<sup>2</sup>The influence of European languages is due to the fact that most of the Arab countries were European colonies during the 19th century.

<sup>3</sup>An other classification is given in [6] where rural and Bedouin Arabic dialects are distinguished because of ethnic and social diversity of Arabic speakers. The author states that Bedouin dialects tend to be more conservative and homogeneous, while urban dialects show more evolutive tendencies.

Arabic in various language levels: Phonological differences between Classical Arabic and spoken Arabic are moderate (compared to other pairs of language-dialect), whereas grammatical differences are the most striking ones. At lexical level, differences are marked with variations in form and with differences of use and meaning.

Indeed, at phonological level, ALG (naturally) shares the most features related to Arabic. In addition to the 28 consonants phonemes of Arabic<sup>4</sup> (given in Table I), ALG consonantal system includes non Arabic phonemes like /g/ as in the word فَاغ (all), and the phonemes /p/ and /v/ used mainly in words borrowed from French like the case of پُومِية (adapted from the French word "pompe" which means a pump) and فَلَيزَة (adapted from the French word "valise" which means a bag). Also, it should be noted that the use of the phonemes (ظ) and (ذ) is very rare, most of the time ظ is pronounced /d'/(ض) and ذ is pronounced /d/(د). The same case is observed for /T/(ث) which is pronounced /t/(ت). Note that the last two substitutions are observed also for Jordanian dialect [9].

TABLE I: Arabic phonemes using SAMPA <sup>5</sup>

Letter	Phoneme	Letter	Phoneme	Letter	Phoneme
أ	/ʔ/	ز	/z/	ق	/q/
ب	/b/	س	/s/	ك	/k/
ت	/t/	ش	/ʃ/	ل	/l/
ث	/T/	ص	/s'/	م	/m/
ج	/Z/	ض	/d'/	ن	/n/
ح	/x/	ط	/t'/	ه	/h/
خ	/X/	ظ	/D'/	و	/w/
د	/d/	ع	/ʔ'/	ج	/j/
ذ	/D/	غ	/G/		
ر	/r/	ف	/f/		
ـَ	/a/	ـِ	/i/	ـُ	/u/
ا	/a:/	ي	/i:/	و	/u:/

Phonological features of ALG will be detailed further in this paper (section V).

#### A. Vocabulary

Algerian dialect has a vocabulary inspired from Arabic but the original words have been altered phonologically, with significant Berber substrates, and many new words and loanwords borrowed from French, Turkish and Spanish. Even though most of this vocabulary is from MSA, there is significant variation in the vocalization in most cases, and the omission or modification of some letters in other cases (mainly the Hamza)<sup>6</sup>. Vocabulary of Algiers's dialect includes verbs, nouns, pronouns and particles. In the following a brief description of each category.

- Verbs

Some verbs in ALG can adopt entirely the same

<sup>4</sup>including three long vowels (ا ā, و w and ي y).

<sup>5</sup>We use the Speech Assessment Methods Phonetic Alphabet for phoneme representation, <http://www.phon.ucl.ac.uk/home/sampa/index.html>.

<sup>6</sup>The Hamza is a letter in the Arabic alphabet, representing the glottal stop.

scheme of MSA verbs by respecting the same vocalization such as in the case of the verb سَمَّى (to name) or سَلَّمَ (to salute). Other verbs are pronounced differently from corresponding MSA verbs by adopting different diacritics marks as the case of the verbs شَرِبَ (to drink). An other set of dialect verbs are obtained by the omission or modification of some letters. In Table II we give some examples of each listed case.

TABLE II: Examples of verbs scheme differences between ALG and MSA.

ALG Verb	Corresponding MSA Verb	Meaning Situation	Situation
سَلَّمَ	سَلَّمَ	To salute	Same scheme
قَابَلَ	قَابَلَ	To confront	Same diacritics marks
شَرِبَ	شَرِبَ	To drink	Same scheme
كَتَبَ	كَتَبَ	To write	/Different Diacritics marks
جَاءَ	جَاءَ	To come	Letters omission or modification
بَقِيَ	بَقِيَ	To remain	
أَكَلَ	أَكَلَ	To eat	
أَتَمَلَ	أَتَمَلَ	To finish	

Another set of ALG verbs are those borrowed from foreign languages especially French such as شَارَجَا which corresponds to the French verb "charger" (to load) or فَأَارَا a modification of the verb "garer" which means to park.

• Nouns

Arabic ALG nouns can be primitive (not derived from any verbal root) or derived from verbs like for verbal names and participles (active and passive), in Table III an exemple is given. We should note that ALG nouns

TABLE III: Example of ALG nouns derived from a verb.

Verb	Verbal name	Acive participle	Passive participle
باع	بيع	بايع	مبيوع
To sale	Sale	Seller	Sold

include an important portion of french words. Most of them are the results of a wide phonological alteration of original words such as موتور ("moteur" in French, motor), لاطونسيون ("la tension", blood pressure) and پوليس ("policier", policeman). Nouns include also numbers which represent units, tens, hundreds, etc. From 1 to 10 the numbers are close to MSA (with different vocalization), except for the numbers 0 and 2: the first one is pronounced as in French /zero/, and the second is زوج, whereas in MSA it is اثنين. From number 11 to 19 the pronunciation in ALG differs from MSA, some letters and diacritics change but the number can be perceived easily by an Arab speaker. Numbers greater than 20 are also close to MSA numbers, only the diacritics marks differ.

• Pronouns

The list of the pronouns is a closed list; it contains

demonstrative and personal pronouns. For relative pronouns, there is only one in Algiers dialect which is أَلِي (that); this pronoun is used for female, masculine, singular and plural. We give in Tables IV and V all ALG used pronouns. It is important to note that the

TABLE IV: Personal pronouns of Algiers dialect.

	Singular		Plural
	Female	Masculine	Female & Masculine
1st Person	أنا I	أنا I	حنا We
2nd Person	أنت You	انت You	انتوما You
3rd Person	هي She	هو He	هوما They

dual in ALG does not exist; there are no equivalent for Arabic pronouns أنتما (second person, dual) and هما (third person, dual). Similarly, personal pronouns relative to feminine plural أنتنّ and هنّ related to second and third person respectively do not exist.

TABLE V: Demonstrative pronouns of Algiers dialect.

	Singular		Plural
	Female	Masculine	Female & Masculine
هادي	هادا	هادو	
This	This	These	
هاديك	هاداك	هادوك	
That	That	Those ones	

• Particles

Particles are used in order to situate facts or objects relatively to time and place. They include different categories such us: prepositions (في in, على on, بـ with), coordinating conjunctions (و and, أو وبعد, after), quantifiers (كل, كلش, قاع, all, شوية, few).

B. Inflection

Algiers dialect is an inflected language such as Arabic. Words in this language are modified to express different grammatical categories such as tense, voice, person, number, and gender. It is well-known that depending on word category, the inflection is called conjugation when it is related to a verb, and declension when it is related to nouns, adjectives or pronouns. We show in the following these linguistic aspects for Algiers dialect.

1) Verbs conjugation: Verb conjugation in ALG is affected (as in MSA) by person (first, second or third person), number (singular or plural), gender (feminine or masculine), tense (past, present or future), and voice (active or passive). Algiers dialect uses as MSA the followings forms:

- The past: Its forms are obtained by adding suffixes relative to number and gender to the verb root and by changing its diacritic marks (see Table VI for a sample).

TABLE VI: The verb كَتَب conjugation in the past tense.

Pronouns		ALG	MSA	English
1st Person	أنا	كُتِبْتُ	كُتِبْتُ	I wrote
	حنا	كُتِبْنَا	كُتِبْنَا	We wrote
2nd Person	أنت	كُتِبْتِ	كُتِبْتِ	You wrote
	أنت	كُتِبْتَ	كُتِبْتَ	You wrote
	أنتوما	كُتِبْتُمَا	كُتِبْتُمَا	You wrote
3rd Person	هي	كُتِبَتْ	كُتِبَتْ	She wrote
	هو	كُتِبَ	كُتِبَ	He wrote
	هوَمَا	كُتِبُوا	كُتِبُوا	They wrote

- The present and future: The present form of a ALG verb is achieved by affixation: the prefixes ت, ن, ي and the suffixes و and ي (Table VIII). The verb could be preceded by the particle راه (in its inflected form<sup>7</sup>) to express a present continuous tense. The future is obtained in the same way as present (same prefixes and affixes) but it must be marked by the ante-position of a particle or an expression that indicates the future like أومبعد (later) or غدوا (tomorrow), next month, ...etc.

TABLE VII: The verb لعب conjugation in the present tense.

Pronouns		ALG	MSA	English
1st Person	أنا	تَلْعَبُ	أَلْعَبُ	I play
	حنا	تَلْعَبُو	تَلْعَبُو	We play
2nd Person	أنت	تَلْعَبِي	تَلْعَبِينَ	You play
	أنت	تَلْعَبُ	تَلْعَبُ	You play
	أنتوما	تَلْعَبُوا	تَلْعَبُونَ	You play
3rd Person	هي	تَلْعَبُ	تَلْعَبُ	She plays
	هو	يَلْعَبُ	يَلْعَبُ	He plays
	هوَمَا	يَلْعَبُوا	يَلْعَبُونَ	They play

- The imperative: It expresses commands or requests, and is used only for the second person. It is generally realised by adding the prefix أ and the suffixes ي and و to the verb.

TABLE VIII: The verb خرج conjugation in the present tense.

Pronouns	ALG	MSA	English
أنت	أُخْرِجِي	أُخْرِجِي	Get out (you, singular, feminine)
أنت	أُخْرِجْ	أُخْرِجْ	Get out (you, singular, masculine)
أنتوما	أُخْرِجُوا	أُخْرِجُوا	Get out (you, plural, feminine & masculine)

2) Declension: Singular word declension in written Arabic corresponds to three cases: the nominative, the genitive, and the accusative which take the short vowels ُ, ِ and َ

respectively attached to the end of the word. These three cases are used to indicate grammatical functions of the words. It should be noted that also the vowels ( ُ, ِ, َ ) represent the *tanween* doubled case endings corresponding to the three cases cited above and express nominal indefiniteness. ALG has dropped these case endings such as all Arabic dialects. The disappearance of final short vowels and dropping of /h/ in certain conditions in many dialects of Arabic are very significant changes [10]. The same author in [8] states: Classical Arabic has three cases in the noun marked by endings; colloquial dialects have none. Thus, a major feature of ALG is that it does not accept the three cases declension of singular nouns and adjectives as written Arabic.

For singular nouns declension to the plural, ALG have the same plural classes as MSA:

- Masculine regular plural: which is formed without modifying the word structure by post-fixing the singular word by ين, unlike written Arabic where the masculine regular plural of a noun is obtained by adding the suffixes ون (for the nominative), and ين (for both the accusative and genitive) depending on the grammatical function of the word. For example, masculine regular plural of MSA word معلم (teacher) could be معلمون (nominative case) or معلمين (accusative or genitive). In contrast, for instance the ALG word رايح (going) always takes رايحين for the regular plural whatever its grammatical category.
- Feminine regular plural: is obtained by adding the suffix ات to the word without changing the structure of the word as in MSA but with a single difference in case endings. Indeed, in MSA, the feminine regular plural has the following marks cases (أَتْ or أَتْ for nominative and أَتْ or أَتْ for accusative and genitive), ALG has only one mark case which is the Sukun السكون (absence of diacritic whose symbol is ). For example the plural of MSA word جميلة is جميلات<sup>8</sup> or جميلات<sup>8</sup> and the plural of ALG word شابة is always شابات (both MSA and ALG words mean beautiful).
- Broken plural: an irregular form of plural which modifies the structure of the singular word to get its plural. As in MSA it has different rules depending on the word pattern. Like singular words, the MSA broken plural takes the three case endings in ALG it does not.

In Table IX we give an example for each ALG plural category.

Another major difference between Algiers dialect and the written Arabic is the absence of the dual (a kind of plural which designs 2 items). Indeed in MSA, for example the dual of ولد (a boy) is designed by وِلْدَان ( the word is post-fixed by َان depending on the case<sup>9</sup>). In ALG Generally, the dual is obtained by the word زوج (two) followed by the plural

<sup>8</sup> جميلات or جميلات also.

<sup>9</sup> ان for nominative case and ين for both accusative and genitive

<sup>7</sup>See next section III-B2

TABLE IX: Examples of ALG plural forms.

Plural	ALG		MSA		English
	Singular	Plural	Singular	Plural	
Regular masculine	فلاح	فلاحين	فلاح	فلاحين/فلاحون	Farmer/Farmers
Case ending	No vowel	ين	ين	ين/ون	
Regular feminine	طبيبة	طبيبات	طبيبة	طبيبات/طبيبات	Doctor/Doctors
Case ending	No vowel	ات	ات	ات/ات	
Irregular	طير	طيور	طير	طيور	Bird/Birds
	يوم	ايامات/ايام	يوم	ايام	Day/Days
Case ending	No vowel	No vowel	ين	ين/ون	

(feminine or masculine) of the noun or the adjective.<sup>10</sup> For example, the dual of *ولد* is *ولدان* (*two boys*)

### C. Syntactic level

1) *Declarative form*: Words order of a declarative sentence in ALG is relatively flexible. Indeed, in common usage ALG sentences could begin with the verb, the subject or even the object. This order is based on the importance given by the speaker to each of these entities; usually the sentence begins with the item that the speaker wishes to highlight. In Table X we give an example of different word orders for a same sentence. It should be noted that the two first forms (SVO,

TABLE X: Example of word order in a ALG declarative sentence.

Order	Dialect Sentence	English
SVO	الولد راح للمسيد	The boy went to school
VSO	راح الولد للمسيد	
OVS	للمسيد الولد راح	
OSV	الولد للمسيد راح	

VSO) are the most used in the every day conversations.

2) *Interrogative form*: In Algiers, any sentence can be turned into a question, in any one of the following ways:

- 1) It may be uttered in an interrogative tone of voice, like *راح تقرا؟* (*Will you revise?*).
- 2) By introducing an interrogative pronoun or particle as *وين راح تقرا؟* (*where will you revise?*).

We list in Table XI the most common interrogative particles and pronouns used in the dialect of Algiers. We mention particularly the particle *ياك* used in questions that accept a yes or no answer.

3) *Negative form*: The particles *ما* and *ماشي* are generally used to express negation. *ما* is used both in Algiers's dialect and MSA, but the form of negation differs between the two languages whereas *ماشي* is specific to the ALG. Using these particles, the negative form is obtained in different ways in ALG (we give in Table XII some examples labeled with each enumerated case):

TABLE XI: Interrogative particles and pronouns in ALG and their equivalents in MSA.

ALG	MSA	English
شكون	من	Who
أما	أى	Which
وين	أين	Where
منين	من أين	From where
واشن / واشن	ماذا	What
باش	بماذا	With what
فاش	في ماذا	In What
وقتاش	متى	When
وعلاش	لماذا	Why
كفاش	كيف	How
شحال	كم	How many

#### • Negation with *ما* particle

- 1) Adding the affixes *ما* and *ش* to conjugated verbs (*ما* as prefix and *ش* as suffix).
- 2) We can enumerate a particular case with the particle *راه* which is equivalent to the verb to be in present tense<sup>11</sup>. The negation is obtained by adding the affixes *ما* and *ش* to the particle *راه* possibly combined with a personal pronoun.

#### • Negation with *ماشي* particle

- 3) The particle *ماشي* can be added at the beginning of a verbal declarative sentence without modification of the sentence.
- 4) The particle *ماشي* can be added at the beginning of a verbal declarative sentence by introducing the relative pronoun *ألى*.
- 5) In the case of a nominal sentence, *ماشي* can be added at the beginning of the sentence by reversing the order of its constituents.
- 6) Also *ماشي* could be added in the middle of a nominal sentence with no modification.

Table XII illustrates some examples of declarative sentences with their negations.

<sup>10</sup> An exception is made for words like *عينين* (two eyes), *ودنين* (two ears), ...

<sup>11</sup>We can not consider this particle as a verb because it could not be conjugated to any other tense

TABLE XII: Declarative sentences with their Negation.

Case	ALG	MSA	English
1	لعبت ما لعبت	لعبت لم تلعب / ما لعبت	she played she didnt play
2	راهي مريضة ما راهيش مريضة	إنها مريضة ليست مريضة	She is ill She is not ill
3	هو ما كتبو ماهي هو ما كتبو	هم كتبوا ليسوا هم من كتبوا	They wrote They are not those who wrote
4	هو ما كتبو ماهي هو ما آلي كتبو	هم كتبوا ليسوا هم الذين كتبوا	They wrote They are not those who wrote
5	الولد مريض ماهي مريض الولد	الولد مريض ليس الولد مريض	The boy is ill The boy is not ill
6	الولد مريض الولد ماهي مريض	الولد مريض الولد ليس مريضا	The boy is ill The boy is not ill

#### IV. CORPUS CREATION

As mentioned above, this work began from scratch. No kind of resources was available for Algiers dialect. The foundation stone of the work was a corpus that we created by transcribing conversations recorded from everyday life and also from some TV shows and movies. This transcription step required conventional writing rules to make the transcribed text homogeneous. Considering the fact that ALG is an Arabic dialect, we adopted the following writing policy: when writing a word in Algiers dialect we look if there is an Arabic word close to this dialect word, if it does exist we adopt the Arabic writing for the dialect word, otherwise the word is written as it is pronounced.

The transcription step produced a corpus of 6400 sentences that we afterwards translated to MSA. Thus, we got a parallel corpus of 6400 aligned sentences. In Table XIII, we give informations about the size of this corpus.

TABLE XIII: Parallel corpus description.

Corpus	#Distinct words	#Words
ALG	8966	38707
MSA	9131	40906

It should be noted that all tasks described above were done by hand. It was time consuming but the result was a clean parallel corpus. Furthermore, ALG side of this corpus has been vocalized with our diacritizer described in [11] and used to develop the first NLP resources dedicated to an Algerian dialect (at our knowledge). The next sections of this paper are dedicated to describe these resources.

#### V. GRAPHEME-TO-PHONEME CONVERSION

As pointed out above, the general purpose of the project TORJMAN is a speech translation system between Modern Standard Arabic and Algiers dialect. Such a system must include a Text-to-Speech module that requires a Grapheme-To-Phoneme converter. We therefore dedicated our efforts to develop this converter by using ALG vocalized corpus described earlier.

Grapheme-to-Phoneme (G2P) conversion or phonetic transcription is the process which converts a written form of a word to its pronunciation form. Grapheme phoneme conversion is not a simple deal, especially for non-transparent languages

like English where a phoneme may be represented by a letter or a group of letters and vice-versa. Unlike English, Arabic is considered a transparent language, in fact the relationship between grapheme and phoneme is one to one, but note that this feature is conditioned by the presence of diacritics. Lack of vocalization generates ambiguity at all levels (lexical, syntactic and semantic) and the phonetic level consequently, such as the word كتب /ktb/, its phonetic transcription could be /kataba/, /kutiba/, /kutubun/, /kutubi/, /katbin/... Algiers dialect obeys to the same rule, without diacritics grapheme-phoneme conversion will be a difficult issue to resolve.

Most works on G2P conversion obey to two approaches: the first one is dictionary-based approach, where a phonetized dictionary contains for each word of the language its correct pronunciation. The G2P conversion is reduced to a lookup of this dictionary. The second approach is rule-based [12], [13], [14], in which the conversion is done by applying phonetic rules, these rules are deduced from phonological and phonetic studies of the considered language or learned on a phonetized corpus using a statistical approach based on significant quantities of data[15], [16]. For Algiers dialect which is a non-resourced language, a dictionary based solution for a G2P converter is not feasible since a phonetized dictionary with a large amount of data is not available. The first intuitive approach (regards to the lack of resource) is a rule based one, but the specificity of Algiers dialect (that we will detail hereafter in the next section.) had led us to a statistical approach in order to consider all features related to this language.

##### A. Issues of G2P conversion for Algiers dialect

Algiers dialect G2P conversion obeys to the same rules as MSA. Indeed, ALG could be considered as a transparent language since alignment between grapheme and phoneme is one to one when the input text is vocalized. But unfortunately, it is not as simple as what has been presented, since ALG contains several borrowed words from foreign languages which most of them have been altered phonologically and adapted to it. Henceforth, the vocabulary of this dialect contains many French words used in everyday conversation. French borrowed words could be divided into two categories: the first includes French words phonologically altered such as the word فاملية (famille in French, family) and the second one includes words which are uttered as in French like the word سور (sûr in French, sure) whose utterance is /syr/(/y/ is not an Arabic phoneme but a French phoneme). This last category constitutes

a serious deal for G2P conversion since these words do not obey to Arabic pronunciation rules.

TABLE XIV: Example of French words used in ALG.

Dialect word	Dialect phonetic transcription	French word	English
كوزينة	/ku:sina/	Cuisine	Kitchen
طابلة	/t'a:bla/	Table	Table
كونكسيون	/konnɛksjɔ̃/	Connection	Connexion
دوفيز	/doviz/	Devise	Currency

In the examples of Table XIV, although the first two words are French, they are phonetized as Arabic words. The French phoneme /t/ is replaced by the Arabic Phoneme /t'/ in the word table. On the other side, the last two words are phoenitized as French words since they are pronounced as in French by Algiers dialect speakers. In order to take account of this word category, the French phonemes like /ɛ/, /ɔ̃/ and /ə/ must be included in Algiers dialect phonemes.

### B. Rule based approach

As stated previously, the rule based approach for G2P conversion applied to ALG requires a diacritized text, that is why we used our ALG vocalized corpus. The diacritized text is converted into its phonetic form by applying the followings rules. It should be noted that most of these rules are those adopted also for Arabic [12], [13] and are applicable only for Arabic words and foreign words phonologically altered in our corpus.

Let consider: BS is a mark of the beginning of a sentence, ES is a mark of the end of a sentence, BL is a blank character, C is a consonant, V is a vowel, LC is a lunar consonant, SC is a solar consonant, and LV is a long vowel. A sample conversion rule could be written as follows:

$$LFT + GR + RGT \Rightarrow /PH/$$

The rule is read as follows: a grapheme GR having as left and right contexts LFT and RGT respectively, is converted to the phoneme PH. Left and right contexts could be a grapheme, a word separator, the beginning or the end of a sentence or empty.

We give in the following all rules that we used for Algiers dialect G2P (the representation of these rules according to the sample below is given in the Appendix (Table XXIII).

- 1) ذ , ظ and ث rules  
In Algiers dialect, the letters ذ , ظ and ث are not used, they are in most cases pronounced as the graphemes د , ض and ت, respectively.
- 2) Foreign letters rules Algiers dialect alphabet corresponds to Arabic alphabet extended to three foreign letters G, V and P.
- 3) Definite article ال
  - The definite article ال is not pronounced when it is followed by a lunar consonant (with does not assimilate the l).  
Example : القمر (the moon)  $\Rightarrow$  /laqmar/  
This rule is the same as in MSA with the

difference that in MSA the ا is pronounced if the definite article is in the beginning of the sentence.

- When the definite article ال is followed by a solar consonant the ل is not pronounced and the consonant following the ل is doubled (gemination).  
Example : السقف (the roof)  $\Rightarrow$  /ʔassqaf/
- When the definite article ال is preceded by a long vowel ي and followed by a solar consonant the definite article is omitted and the solar consonant is doubled (gemination).  
Example : في الدار  $\Rightarrow$  فدار  $\Rightarrow$  /fddAr/

- 4) Words Case-ending  
Words case ending in Algiers dialect is the Sukun (Absence of diacritics), so the last consonant of a word should be pronounced without any diacritic.  
Example : قبل (before)  $\Rightarrow$  /qbal/
- 5) Long vowel rules  
When ا , و and ي appear in a word preceded by the short vowels َ , ُ and ِ , respectively, their relative long vowels are generated.  
Examples:  
كأس (a cup)  $\Rightarrow$  /ka:s/  
فول (beans)  $\Rightarrow$  /fu:l/  
كبير (a well)  $\Rightarrow$  /kbi:r/
- 6) Glottal stop rule  
In Algiers dialect, when a word begins with a Hamza, its phonetic representation begins with a glottal stop. in the end of a word the Hamza preceded by ا is not pronounced.  
Example: أسكت (stop talking)  $\Rightarrow$  /ʔaskut/ and سماء (sky)  $\Rightarrow$  /smʔ/  
It should be noted that the Hamza in the middle of the word is replaced by the long vowels ا or ي in Algiers dialect. For example the Arabic words بئر (hole) and فأس (poleax) correspond to /bi:r/ and /fa:s/, respectively.
- 7) Alif Maqsura rule ي  
Alif Maqsura ي (which is always preceded by a fatha) at the end of a word is realized as the short vowel /a/.  
Example: رمى (he throws)  $\Rightarrow$  /rmaa/
- 8) Alif Madda آ  
Alif madda آ is realized as alef /ʔ/ with the long vowel /a:/.  
Example: آمن (he trusts)  $\Rightarrow$  /ʔa:man/
- 9) Words ending with ة  
The ة is not pronounced in Algiers dialect unlike in MSA where it is realized with the two phonemes /t/ and /h/ (depending on the word position)  
Example: طفلة (a girl)  $\Rightarrow$  /t'afla/
- 10) Words ending with ه



The  $\text{و}$  is not pronounced in Algiers dialect when it is preceded by  $\text{ت}$ .

Example: كتابه (his book)  $\implies$  /kta:bu/

- 11) Words containing the sequences  $\text{ب, ن}$

When a  $\text{ن}$  is followed by a  $\text{ب}$ , the  $\text{ن}$  is pronounced as /m/

Example: منبر (a foretop)  $\implies$  /mambar/

- 12) Gemination rule

When the Shadda appears on a consonant, this consonant is doubled (geminated)

Example: سكر (sugar)  $\implies$  /sukkur/

It should be noted that most of these rules could be applied for other Algerian dialects and Arabic dialect close to them such Tunisian and Moroccan.

**Experiment:** As indicated above for experiment we used our ALG vocalized corpus which includes three categories of words:

- 1) Arabic words.
- 2) French words phonologically altered and their pronunciation is realized with Arabic phonemes.
- 3) French words for which the pronunciation is realized with French phonemes.

We applied phonetization rules seen below on the ALG corpus. In addition to Arabic words, French words of the second category are correctly phonetized because their phonetic realization is close to Algiers dialect. For example the word كوزينة (kitchen, original French word is cuisine) which is a borrowed French word phonologically altered is correctly converted as /ku:zina/, while a word in the third category as كونكسيون (connection, original French word connexion) is incorrectly converted to /ku:niksju:n/ since it is realized /kɔnnɛksjɔ̃/ with French phonemes. Considering these words, system accuracy is 92%. The issue of these words is that we can not introduce rules for French words written in Arabic script, since the relation between Arabic graphemes and French phonemes is not one to one. For example the graphemes  $\text{و}$  in a French word written in Arabic script could correspond to the French phonemes /y/, /u/, /ɔ/ or /O/ (see some examples in Table XV).

TABLE XV: Examples of mappings between Arabic grapheme  $\text{و}$  and French phonemes.

Dialect word	French phonetic transcription	French word	English
سور	syR	Sûre	Sure
پور	pɔR	Port	Port
سودور	sudœR	Soudeur	Wilder

### C. Statistical Approach

Rule based approach adopted above does not take into account French words used in ALG which are pronounced as in French language. This issue takes us to choose a statistical approach in order to consider this feature. We use statistical machine translation system where source language is a text (a set of graphemes) and target language is its phonetic

representation (a set of phonemes). This system uses Moses package[17], Giza++[18] for alignment and SRILM[19] for language model training. The main motivation of using a statistical approach is that we can include French phonemes in the training data. For building this system, the first component is a parallel corpus including a text and its phonetic representation. Actually, this resource is not available, so we created it by using the rule based converter described above. We proceed as follows: we used the rule based system to convert Arabic words and French words phonologically altered (category 1 and 2) to Arabic phonemes. Whereas for French words realized with French phonemes (category 3), we began by identifying them and we transliterated them to their original form in Latin script, then converted them to French phonemes (using a free French G2P converter), all these operations were done by hand. For example the word كونكسيون is transliterated to connexion then converted to /kɔnnɛksjɔ̃/.

This system operates at grapheme and phoneme level, we split the parallel corpus into individual graphemes and phonemes including a special character as word separator in order to restore the word after conversion process (see Table XVI).

TABLE XVI: Examples of aligned graphemes and phonemes.

و	ت	ـ	ك	و	م	ـ	ن
Null	/t/	/u/	/k/	Null	/s/	/a/	/n/

**Experiment:** For evaluating the statistical approach, we split the parallel corpus into three datasets: training data (80%) tuning data (10%) and testing data (10%). First we tested the statistical approach on a corpus containing only Arabic words and French words phonologically altered (category 1 and 2). We got an accuracy of 93%. Then we proceeded to a test on a corpus including the three words categories, system accuracy decreases to 85%. This result is due to the increase of hypothesis number of each grapheme because of introducing French phonemes in the training data. The graphemes  $\text{و}$  for example in some Arabic words (category 1) are phonetized as the French phonemes /y/ or /ɔ/ instead of the Arabic long vowel /u:/, the phoneme /ɔ/ instead of /u:n/. Contrary to that some words in category 3 are phonetized with Arabic phonemes by substituting for example the phonemes /y/, /u/, /ɔ/ or /O/ by the /u:/, and /ɛ/ by /a:/.

### D. Discussion

At first glance, and regards to accuracy rates, we could deduce that rule based approach is more efficient than statistical approach (92% vs 85%). Rule based approach does not take into account French words of category 3, it achieves efficient results only for Arabic words and French phonologically altered words (category 1 and 2). Results of statistical approach must be analysed regards to the small amount of the training data. On another side, a hybrid approach could be adopted: instead of using one corpus including all categories of words for training the statistical G2P converter, we can use two corpora: the first one including words of categories 1 and 2, could be processed by rule based approach. The second corpus is a parallel corpus including words of category 3 with their French phonetization used for training the statistical G2P

converter. Unfortunately, we have not sufficient data for testing such a converter, since our corpus includes only about 1k words of category 3. In terms of resources, this work allowed us to build a phonetized dictionary for Algiers dialect; at our knowledge no such resource is available at this time.

## VI. MORPHOLOGICAL ANALYZER FOR ALGERIAN DIALECT

### A. Related works

Compared to MSA, there are a little number of Morphological Analysers (MA) dedicated to Arabic dialects. Works in this area could be divided into two categories. The first one includes MA that are built from scratch such as in [20] and [21], the second includes works that attempt to adapt existing MSA Morphological Analysers to Arabic dialect. This trend is adopted for several dialects since it is not time consuming. In [22], authors used BAMA Buckwalter Arabic Morphological Analyser [3] by extending its affixes table with Levantine/Egyptian dialectal affixes. The same approach is adopted in [23] where a list of dialectal affixes (belonging to four Arabic dialects) was added to Al-Khalil [24] affix list. Authors in [25] converted the ECAL (Egyptian Colloquial Arabic Lexicon) to SAMA (Standard Modern Arabic Analyser) representation [26]. For Tunisian dialect, authors in [27] adapted Al-Khalil MA, they create a lexicon by converting MSA patterns to Tunisian dialect patterns and then extracting specific roots and patterns from a training corpus that they created.

### B. Adopted Approach

To build a MA for Algiers dialect, we decide to adapt BAMA, since it does not consume time and takes profit from the fact that it is widely used. BAMA is based on a dictionary of three tables containing Arabic stems, suffixes and prefixes and three compatibility tables defining relations between stems, prefixes and suffixes. Adaptation of BAMA is got by populating these tables by dialect data.

### C. Building the dialect dictionary

We built dialect dictionary by adopting the following principle: in order to exploit BAMA dictionary, we kept from it all entries that belong also to ALG with some modification (for example MSA prefixes  $\text{بـ}$ ,  $\text{تـ}$  and  $\text{لـ}$  are used in ALG so we kept them as ALG prefixes). Beside that, we deleted all entries which are not suitable for Algiers dialect. Moreover, we created entries that are purely dialectal and which did never exist in MSA dictionary.

1) *Affixes tables*: For affixes tables, common affixes between MSA and ALG are kept (in prefixes and suffixes tables), whereas all other MSA affixes which do not belong to dialect were deleted. However, some dialect affixes which do not exist in MSA were added to affixes tables. Note that when an affix is deleted, all complex affixes where it occurs are also deleted.

- 1) Prefixes table: We kept some prefixes unchanged like prefixes  $\text{بـ}$  and  $\text{تـ}$  that precede imperfect verbs (for the singular third person masculine and feminine, respectively). We eliminated purely MSA prefixes<sup>12</sup>

and all complex prefixes where they appear instead of the prefix  $\text{سـ}$  (expressing the future when it precedes imperfect verbs) and the prefix  $\text{فـ}$ <sup>13</sup> (conjunction), some examples are given in Table XVII.

TABLE XVII: Examples of kept, deleted and added prefixes in ALG prefixes table.

Kept pref.	Description
$\text{تـ}$ , $\text{بـ}$	Imperfect Verb Prefix (sing., third person, masc., fem.)
$\text{الـ}$	Noun Prefix (definite article)
$\text{لـ}$ , $\text{بـ}$	Preposition Prefix
Del. pref.	Description
$\text{فـ}$	Conjunction Prefix
$\text{سـ}$	Future Imperfect Verb Prefix
$\text{فـ}$ , $\text{بـ}$	Conj.Pre.+Preposition Pre.+Definite Art. Pre.
Add. pref.	Description
$\text{فـ}$	Preposition Prefix
$\text{فالـ}$	Preposition Pre.+Definite Art. Pre.
$\text{ينـ}$	Perfect verb pre. (past voice, (sing., masc.) and (plu, masc/fem.))
$\text{تنـ}$	Perfect verb pre. (past voice, (sing. fem.))

- 2) Suffixes table: We also eliminated all MSA suffixes not used in Algiers dialect mainly:
  - Suffixes related to the dual both feminine and masculine,
  - Feminine plural suffixes,
  - All word case endings suffixes

All complex suffixes where they appear were also deleted. Likewise, we added dialectal suffixes like the suffix  $\text{شـ}$  for negation and all complex suffixes that must be included with it.

We integrated also a set of suffixes to take into account all various writings of dialects words which are not normalized. An example is the suffix  $\text{و}$ , which could express the plural (feminine and masculine) in the end of a verb, a possessive pronoun at the end of a noun exactly like the MSA suffix  $\text{هـ}$ . We give in table XVIII a set of examples of each case.

2) *Stems table*: Dialect stems table was populated by the lexicon of Algiers dialect corpus and MSA stems included in BAMA. We used a part (85%, 9170 distinct words) of our ALG corpus for creating dialect stems, the remaining 15% (1618 distinct words) is used for test.

### Stems from ALG corpus lexicon

First, we began by extracting a list of nouns easily identifiable by affixes  $\text{ة}$  and definite article  $\text{الـ}$  (used only with nouns). We deleted these two affixes from all extracted words, then from obtained list of words we created stem entries according to BAMA. Next, the rest of the corpus was analysed and classified into three sets: function words, verbs and nouns (which do not include  $\text{ة}$  and  $\text{الـ}$  suffixes) and converted to stems according to BAMA stems categories. Let us indicate that we added some stems categories to take into account all dialectal features. For example, in MSA the perfect verb stem category

<sup>12</sup>Prefixes that could not belong to Algiers dialect.

<sup>13</sup>Note that  $\text{فـ}$  as MSA conjunction prefix has been deleted (since it does not exist in ALG), and  $\text{فـ}$  as preposition prefix has been created.

TABLE XVIII: Examples of kept, deleted and added suffixes in ALG suffixes table.

Kept Aff.	Description
ين	Accusative/genitive noun Suffix(masc.,plu.)
ات	Noun Suffix(fem.,plu.)
ت	perfect verb suffix (fem.,sing)
Del. suff.	Description
ن	Perfect/Imperfect Verb Suffix(subject, plu., fem.)
تما	Perfect/Imperfect Verb Suffix(subject, dual., fem/masc., 2nd person)
هما	Perfect/Imperfect Verb Suffix(direct object, dual., fem/masc., 3rd person)
ون	Nominative Noun Suffix (masc.,plu.)
ان	Nominative Noun Suffix (masc.,dual)
هن	Perfect/Imperfect Verb Suffix(direct object, plu., fem.)
تهن	Perfect Verb Suffix(subject sing.,2nd person,masc.,direct object, plu., 3rd person, fem.)
Add. suff.	Description
ش	Perfect/Imperfect Verb Negation Suffix
همش	Perfect/Imperfect Verb Negation Suffix (direct object,plu., 3rd person, masc./fem)
كمش	Perfect/Imperfect Verb Negation Suffix (direct object,plu., 2nd person, masc./fem.)
و	Per./Imp. Verb Suffix(direct object,plu.,masc.,fem.)

with the pattern **فَعَلَ** covers the three persons, the two genders, the single, the dual and plural; just relative suffixes are added to it to have its different inflected forms. In ALG, we split this stem category into two distinct stems: **فَعَلَ** and **فَعِلْ** to cover all perfect verbs inflected forms, in Table XIX we give an example related to the stem **سَمِعَ** (to hear).

TABLE XIX: Example of splitting a MSA stem to two Dialectal stems.

Eng. pro.	Dia. pro.	Dia. verb	Dia. stem	MSA pro.	MSA verb	MSA stem
She	هي	سَمِعَتْ	سَمِعَ	هي	سَمِعَتْ	سَمِعَ
They	هوَمَا	سَمِعُوا		هم	سَمِعُوا	
He	هو	سَمِعَ	هو	سَمِعَ		
We	هَنَا	سَمِعْنَا	نحن	سَمِعْنَا		

Exploiting MSA BAMA stems

1) Verbs

The main idea for creating ALG verb stems from MSA stems is using verbs pattern. For example the verbs having ALG pattern **فَعَلَ** are in most cases

Arabic verbs with the patterns **فَعَلَ**, **فَعِلْ** or **فَعِلْ**. Some other ALG verbs keep the same pattern as in MSA like verbs with the patterns **فَعَلَ**

From stems table, we extracted all perfect verbs having the patterns **فَعَلَ**, **فَعِلْ** and **فَعِلْ**. After that, the verbs having the three first patterns are converted to Algiers dialect pattern by changing diacritic marks to **فَعَلَ** while the verbs corresponding to pattern **فَعَلَ** are kept as they are (since this pattern is used in ALG). At this stage, we constructed a set of Arabic verb stems having dialect pattern, we analysed them and eliminated all stems that are not used in ALG. We give in Table XX some examples.

We proceed as explained above for other patterns as **فَعَلَ**, **فَعِلْ**, **فَعِلْ**. It should be noted that,

TABLE XX: Examples of converted stems from MSA to ALG.

Stems	ALG Dialect	MSA	English
ضرب	ضَرَبْ	ضَرَبَ	He beat
شرب	شَرَبْ	شَرَبَ	He drunk
بدل	بَدَّلْ	بَدَّلَ	He changed
كبر	كَبَّرْ	كَبَّرَ	He grew

we constructed imperfect verb stems and command verb stems from the ALG perfect verb stems that we created as described above.

2) Nouns

We kept all proper nouns from MSA stems table because it contains an important number of entries related to countries, currencies, personal nouns,... We analysed all other types of words and kept from them those existing in ALG by modifying diacritics, adding or deleting one or more letters.

3) Function words

We deleted all function words that do not exist in ALG like relative pronouns and personal pronouns related to the dual and feminine plural, then we translated remaining ones to ALG.

Note that we introduced dialect stems with non Arabic letters **ف** *G*, **و** *V*, and **پ** *P* in stems table and we modified BAMA code to consider words containing these letters. Also, since every stem entry in BAMA contains an English glossary, when creating a dialect entry, we added the Arabic word to English glossary, so for each dialect entry is associated an English and Arabic glossary.

After creating affixes and stems tables for ALG, compatibility tables of BAMA were updated according to the data included in these tables.

D. Experiment

As mentioned above, we tested our MA on the Algiers Dialect corpus, the test set contains 1618 distinct words extracted from 600 sentences chosen randomly. We consider

that a word is correctly analysed if it is correctly decomposed to prefix+stem+suffix and if all the features related to them are correct (POS, gender, number, person). We first began by testing the MA with stems extracted only from the ALG corpus lexicon, then we introduced stems created from the MSA stems table. We list in Table XXI the obtained results.

TABLE XXI: Results of ALG morphological Analyser.

Results	ALG corpus stems	MSA stems+ALG corpus stems
# Analysed words	703	1115
Percentage	43%	69%
# Unanalysed words	915	503
Percentage	57%	31%

We examined the words for which no answer were given by the morphological analyzer(see Table XXII), most of the cases are:

- French words which do not exist in the stem table like تريسييتي (électricité , electricity), or words like انجنيور (ingénieur, engineer) and النيمرو (numéro, number) that are included in stems table but with an other orthography (respectively انجينيور and النيمرو). The same case is observed for nouns written with long vowel ا in the end instead of ة such as پلاسا (place).
- We noticed also that some words are written with missed letters like the word النساء which appears in stems table as النساء. The same case is noticed for قتلو or قال لي or قالى (he said to me) instead of قتلو or قلت لو (I said to him) instead of قتلو or قلت لو.
- Some Unanalyzed words also are proper nouns.

TABLE XXII: Examples of unanalyzed words.

Unanalyzed word	Corresponding stem	English
انترنت	انترنت	Internet
امبعد	اومبعد	After
تريماستر	تريمستر	Trimester
تليفون	تليفون	Phone

## VII. CONCLUSION

This paper summarize a first attempt to work on Algerian Arabic dialects which are non-resourced languages. These dialects lag behind compared to other dialects of the Middle-east for which several works were dedicated and produced many NLP tools. The presented work is the first part of a big project of Speech translation between MSA and Algerian dialects. We focus in this first part on the one spoken in Algiers and its periphery. We began by a study showing all features related to it, then we introduced resources that we created from scratch. This process was expensive in terms of time and human effort but the results were worth it. We get a cleaned corpus of Algiers dialect aligned to MSA, this corpus is the first parallel corpus which includes Algerian dialect to date. We presented also the Grapheme-to-Phoneme converter that we created for Algiers dialect. We combined a rule based approach to a statistical approach. The level of correctness for

the G2P converter is about 85%. In terms of corpus resources, this task enabled us to transcribe the ALG corpus to a phonetic form. We also proposed a morphological analyser for AIG that we adapted from the well known BAMA dedicated for MSA. We reached an accuracy rate of 69% when evaluating it on a dataset extracted from ALG corpus. Our future work before developing a statistical machine translation system, is to extend the corpus we created to other Algerian Arabic dialects, and to adapt all tools dedicated to ALG to these dialects.

## ACKNOWLEDGEMENT

This work has been supported by PNR (Projet National de Recherche of Algerian Ministry of Higher Education and Scientific Research).

## REFERENCES

- [1] S. Harrat, K. Meftouh, M. Abbas, and K. Smaili, "Building resources for algerian arabic dialects," in *Proceedings of Interspeech*, 2014, pp. 2123–2127.
- [2] —, "Grapheme to phoneme conversion: An arabic dialect case," in *Proceedings of 4th International Workshop On Spoken Language Technologies For Under-resourced Languages SLTU*, 2014, pp. 257–262.
- [3] B. Tim, "Buckwalter arabic morphological analyzer version 1.0," *Linguistic Data Consortium LDC2002L49*, 2002.
- [4] K. Kirchhoff, J. Bilmes, S. Das, N. Duta, M. Egan, G. Ji, F. He, J. Henderson, D. Liu, M. Noamany, P. Schone, R. Schwartz, and D. Vergyi, "Novel approaches to arabic speech recognition: Report from the 2002 johns-hopkins summer workshop," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '03)*, vol. 1, April 2003, pp. 1–344–1–347.
- [5] R. Hetzron, *The Semitic Languages*, ser. Routledge language family descriptions. Routledge, 1997. [Online]. Available: <https://books.google.com/books?id=nbUOAAAQAAJ>
- [6] J. C. Watson, *The phonology and morphology of Arabic*. Oxford university press, 2007.
- [7] A. Boucherit, *L'Arabe parlé à Alger*. ANEP Edition, 2002.
- [8] C. A. Ferguson, "Diglossia," *Word*, vol. 15, pp. 325–340, 1959.
- [9] F. H. Amer, B. A. Adaileh, and B. A. Rakhieh, "Arabic diglossia: A phonological study," *Argumentum 7, Debreceni Egyetem Kiadó, Tanulmány*, pp. 19–36, 2011.
- [10] C. A. Ferguson, "Two problems in arabic phonology," *Word*, vol. 13, pp. 460–478, 1957.
- [11] S. Harrat, M. Abbas, K. Meftouh, and K. Smaili, "Diacritics restoration for arabic dialect texts," in *Proceedings of Interspeech*, 2013, pp. 125–132.
- [12] M. Alghamdi, H. Almuhtasab, and M. Alshafi, "Arabic phonological rules," *Journal of King Saud University: Computer Sciences and Information (in Arabic)*, vol. 16, pp. 1–25, 2004.
- [13] Y. A. El-Imam, "Phonetization of arabic: rules and algorithms," *Computer Speech Language*, vol. 18, no. 4, pp. 339–373, 2004.
- [14] M. Zeki, O. O. Khalifa, and A. Naji, "Development of an arabic text-to-speech system," in *International Conference on Computer and Communication Engineering (ICCCE)*. IEEE, 2010, pp. 1–5.
- [15] P. Taylor, "Hidden markov model for grapheme to phoneme conversion," in *Proceedings of Interspeech*, 2005, pp. 1973–1976.
- [16] K. U. Ogbureke, P. Cahill, and J. Carson-Bermdsen, "Hidden markov models with context-sensitive observations for grapheme-to-phoneme conversion," in *Proceedings of Interspeech*, 2010, pp. 1105–1108.
- [17] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," *Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstration session*, pp. 177–180, 2007.

- [18] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, volume 29, number 1, pp. 19–51, 2003.
- [19] A. Stolcke, "Srlm – an Extensible Language Modeling Toolkit," in *Proceedings of Interspeech*, Denver, USA, 2002, pp. 901–904.
- [20] N. Habash and O. Rambow, "Magead: A morphological analyzer and generator for the arabic dialects," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 681–688.
- [21] M. Altantawy, N. Habash, and O. Rambow, "Fast yet rich morphological analysis," in *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 116–124.
- [22] W. Salloum and N. Habash, "Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation," in *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*. Association for Computational Linguistics, 2011, pp. 10–21.
- [23] K. Almeman and M. Lee, "Towards developing a multi-dialect morphological analyser for arabic," in *4th International Conference on Arabic Language Processing*, 2012, pp. 19–25.
- [24] A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane, M. O. A. O. Bebah, and M. Shoul, "Alkhalil morpho sys: A morphosyntactic analysis system for arabic texts," in *Proceedings of 7th International Computing Conference in Arab ACIT*, 2011.
- [25] N. Habash, R. Eskander, and A. Hawwari, "Morphological analyzer for egyptian arabic," in *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology SIGMORPHON*. Association for Computational Linguistics, 2012, pp. 1–9.
- [26] D. Graff, M. Maamouri, B. Bouziri, S. Krouna, S. Kulick, and T. Buckwalter, "Standard arabic morphological analyzer (SAMA) version 3.1," *Linguistic Data Consortium LDC2009E73*, 2009.
- [27] I. Zribi, M. E. Khemakhem, and L. H. Belguith, "Morphological analysis of tunisian dialect," in *International Joint Conference on Natural Language Processing*, 2013, pp. 992–996.

## Appendix

TABLE XXIII: Algiers dialect Rules for G2P conversion.

#	Rule title	Rule
1	ذ, ظ and ث rules	$\{C, V\} + \text{ذ} + \{C, V\} \Rightarrow /d/$
		$\{C, V\} + \text{ظ} + \{C, V\} \Rightarrow /d'/$
		$\{C, V\} + \text{ث} + \{C, V\} \Rightarrow /T/$
2	Foreign letters rules	$\{C, V\} + \text{ف} + \{C, V\} \Rightarrow /g/$
		$\{C, V\} + \text{و} + \{C, V\} \Rightarrow /v/$
		$\{C, V\} + \text{پ} + \{C, V\} \Rightarrow /p/$
3	Definite article ال	$\{LC\} + \text{ال} + \{BL, BS\} \Rightarrow /l/ + /LC/$
		$\{SC\} + \text{ال} + \{BL - BS\} \Rightarrow /?al/ + /SC/ + /SC/$
4	Words Case-ending	$\{BL, ES\} + C + \{C, V\} \Rightarrow /C/$
5	Long vowel rules	$\{C + \text{ا}\} + \text{ـ} + \{C\} \Rightarrow /a : /$
		$\{C + \text{و}\} + \text{ـ} + \{C\} \Rightarrow /u : /$
		$\{C + \text{ى}\} + \text{ـ} + \{C\} \Rightarrow /i : /$
6	Glottal stop rule	$\{C, V\} + \text{أ} + \{BS, BL\} \Rightarrow /?/$
		$\{BL, ES\} + \text{ء} + \{ \text{ا} \} \Rightarrow /Null/$
7	Alif Maqsura rule ع	$\{BL, ES\} + \text{ع} + \{ \text{ـ} + C \} \Rightarrow /a/$
8	Alif Madda آ	$\{C\} + \text{آ} + \{C\} \Rightarrow /?a : /$
9	Words ending with ð	$\{BL, ES\} + \text{ð} + \{C, V\} \Rightarrow /Null/$
10	Words ending with ɛ	$\{BL, ES\} + \text{ɛ} + \{ \text{ـ} \} \Rightarrow /Null/$
11	Words containing the sequences ب,ن	$\{ \text{ب} \} + \text{ن} + \{C, V\} \Rightarrow /m/$
12	Gemination rule	$\{V\} + \text{و} + \{C\} \Rightarrow /CC/$

# An Enhanced Automated Test Item Creation Based on Learners Preferred Concept Space

Mohammad AL-Smadi

Computer Science Department  
Jordan University of Science and Technology  
P.O.Box: 3030 Irbid 22110, Jordan

Margit Höfler

Institute of Psychology  
University of Graz  
Graz, Austria

Christian Gütl

Graz University of Technology  
Curtin University of Technology, Perth,  
WA Graz, Austria

**Abstract**—Recently, research has become increasingly interested in developing tools that are able to automatically create test items out of text-based learning contents. Such tools might not only support instructors in creating tests or exams but also learners in self-assessing their learning progress. This paper presents an enhanced automatic question-creation tool (EAQC) that has been recently developed. EAQC extracts the most important key phrases (concepts) out of a textual learning content and automatically creates test items based on these concepts. Moreover, this paper discusses two studies for the evaluation of EAQC application in real learning settings. The first study showed that concepts extracted by the EAQC often but not always reflect the concepts extracted by learners. Learners typically extracted fewer concepts than the EAQC and there was a great inter-individual variation between learners with regard to which concepts they experienced as relevant. Accordingly, the second study investigated whether the functionality of the EAQC can be improved in a way that valid test items are created if the tool was fed with concepts provided by learners. The results showed that the quality of semi-automated creation of test items were satisfactory. Moreover, this depicts the EAQC flexibility in adapting its workflow to the individual needs of the learners.

**Keywords**—Automated Assessment; Automatic Test-Item Creation; Self-Regulated Learning; Evaluation of CAL systems; Pedagogical issues; Natural-Language Processing

## I. INTRODUCTION

The ability of learners to self-regulate their learning process is a key competence in life-long learning. One efficient way of such self-regulation is to monitor and assess their learning progress by self-assessment (e.g., [1], [2], [3]). Self-assessment supports learners to focus on the most important aspects of the material. Moreover, it helps them to increase their involvement in the learning process [1]. However, learners often face problems if they have to generate questions on their own. For instance, in [4] medical students had to generate questions in order to enhance their metacognition strategies such as self-regulation and problem-solving. The results showed that some of the students had difficulties to create high-order questions (i.e., questions that ask for synthesis and evaluation of the information rather than for simple recall). Furthermore, even if the students had the opportunity to enhance their poorly formulated questions, they needed appropriate guidance to do so. Hence, it might be necessary to train learners in strategies and procedures of self-questioning to increase their competence in question generation [2].

But how can learners receive adequate support in creating good questions? Previous research has presented several guidelines and strategies to support the generation of effective questions (e.g., [5], [6], [7]). For instance, the Taxonomy of Bloom ([8], [9]) provides a powerful framework of how to build questions tailored to different learning goals and levels of cognitive processing (e.g., on a basic level, learners should be able to understand a formula whereas on a more sophisticated level they should be able to apply or explain it). Furthermore, [10] presented examples of guideline questions which might be useful for self-assessment:

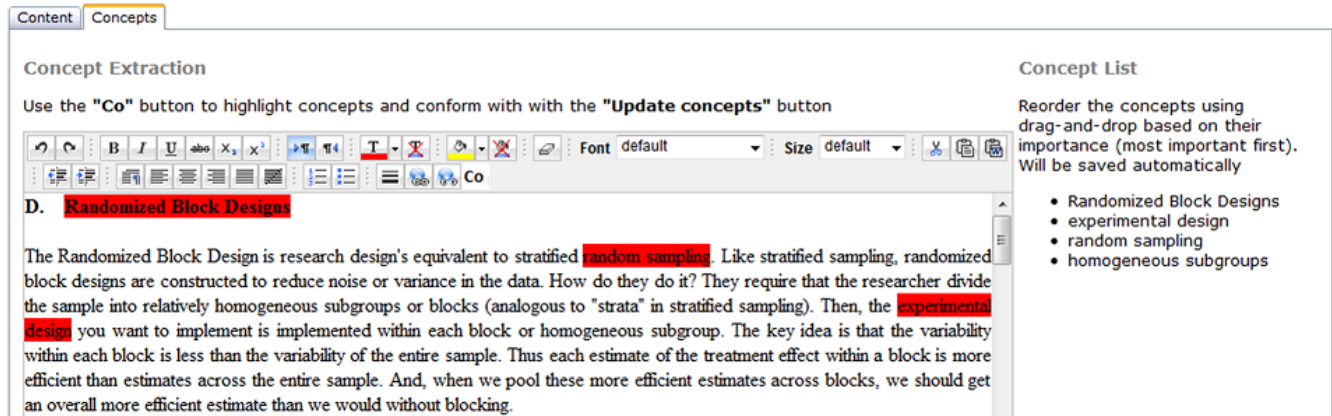
- 1) Can I summarize the main idea of the text?
- 2) Can I list the five most important learning points of the chapter?
- 3) Can I write a short comment?
- 4) Can I discuss the topic raised in this chapter?
- 5) Are the important learning points I list consistent with those proposed by my classmates and teacher?

What the questions presented above have in common is that they rather address the general ideas and themes of a learning content than its details (e.g., there is no need to know every single detail about a topic if someone has to comment on it). In fact, Bugg and McDaniel [11] showed that conceptual questions addressed to the gist of the text (which required the integration of information across sentences) led to a better performance on a memory test compared to questions that required detailed knowledge about the text. Hence, if questions are generated from a learning content, it might be more valuable to focus on the most important keywords or key phrases than on the details.

According to [12], keywords are defined as a sequence of one or more words and provide a compact description of a document's content whereas key phrases consist of two or more key words and named identities (p.140). In general, key words and key phrases (in the following subsumed to the term concepts) help readers quickly to identify whether a text might be relevant to their needs or not (e.g. [13]). For example, some concepts of the current text (as reflected in the key words section of this paper) are Automated Assessment, and Self-Regulated Learning. Using these concepts, the readers of this journal can quickly identify whether this article might be relevant to their research or not. Likewise, in the context of learning, concepts provide an overview about the learning content and questions that are based on such concepts may



## Explore subjects



The screenshot displays the 'Explore subjects' interface. At the top, there are tabs for 'Content' and 'Concepts'. Below this is the 'Concept Extraction' section, which includes instructions: 'Use the "Co" button to highlight concepts and conform with with the "Update concepts" button'. A rich text editor toolbar is visible, with the 'Co' button highlighted. The main text area shows a section titled 'D. Randomized Block Designs' with several words highlighted in red: 'random sampling', 'experimental design', and 'overall more efficient estimate'. To the right, a 'Concept List' is shown, containing a bulleted list of concepts: 'Randomized Block Designs', 'experimental design', 'random sampling', and 'homogeneous subgroups'. Instructions for reordering concepts by importance are also present.

Fig. 1: Screenshot with the learner view form the enhanced EAQC for manual selection of concepts as part of the automated creation of test items.

guarantee that the major themes are covered. However, it is typically left up to the learner to define which concepts reflect a topic best. This suggests that learners might fail to create appropriate questions for self-assessment activities because of inappropriately chosen concepts and the lack of required knowledge and/or skills.

Another approach to support learners in creating questions subsumes software tools that are capable to create questions automatically from a given (text-based) learning content. In recent years, research has become more and more interested in developing such tools ( e.g., [14], [10], [15], see also [16], for an overview). However, generating questions automatically from a given text is still a challenge. For instance, such questions often lack meaningfulness, an adequate level of difficulty, and appropriate answers or even relevance (see e.g. [17] , [18], [19]).

### A. Problem Statement

Automated question-creation tools typically generate only one type of questions (e.g. open-ended questions that ask for a free answer or multiple-choice questions that ask to find the correct answer among a set of provided distractors). For self-directed learning, having a flexible automated assessment tool that provides different types of questions and preferably also the correct answers has become a need. In addition, these questions should be based on the most relevant concepts of the learning material in order to be effective. However, this research is still active with room for a lot of improvements regarding how the most relevant concepts from natural language texts can be identified (e.g., [12], [13], [20], [21]; a discussion of concept extraction methods can be found in [22] or [23]). Moreover, an effective automated question-creation tool might not only provide questions which reflect the most important concepts of a text best but even allow learners based on their personal needs to determine on which concepts of the text the questions should encounter.

Recently, [24], [25] have presented an enhanced automated question creator (EAQC) tool. EAQC is able to create four types of test items out of English or German learning content.

The provided test item types are open-ended items (which require a free answer), single-choice items (in which a given statement is true or false), multiple-choice items (in which one correct answer has to be found within a given number of distractors), and completion exercises (in which one key concept is missing in a statement). Moreover, the authors have conducted an evaluation study in order to investigate the quality of the test items created by EAQC. To this end, they had students rate the test items with regard to different quality criteria adapted from [26]. In addition to pertinence and terminology of test items, these quality criteria evaluate the quality of the answers (open-ended questions and completion exercises) and distractors (multiple-choice items). Preliminary results from the evaluation study [24] showed that EAQC-based test items did not differ in pertinence and level from manually (by an instructor) created test items and also the provided answers were qualitatively on a par with their hand-made counterparts. However, the terminology of some of the test items and the quality of the distractors for the multiple-choice items were rated as rather poor by the participants. Nevertheless, the overall quality of the test items created by the EAQC was satisfactory in a way that the test items were assumed to be of use for self-regulated learning or even in real test settings (see [27]).

Test items generated by EAQC are based on concepts that were extracted out of the learning content. Due to this process, the resulting questions are expected to cover the main aspects of the learning content. However, it is not clear whether all automatically extracted concepts are, at least from the viewpoint of an individual learner, important. In such case, the resulting automatically created test items might cover aspects that are not supportive for the individual learner at all. Previous research [24], [27] showed that learners experienced that not all concepts automatically extracted by EAQC were relevant [24]. This is a first indication that automatically extracted concepts might not always reflect the learner's view.

To the best of our knowledge, there is no study that directly addresses the question of whether the concepts extracted by an automated approach match the same concepts a learner would extract. Hence, it would be of avail if the EAQC

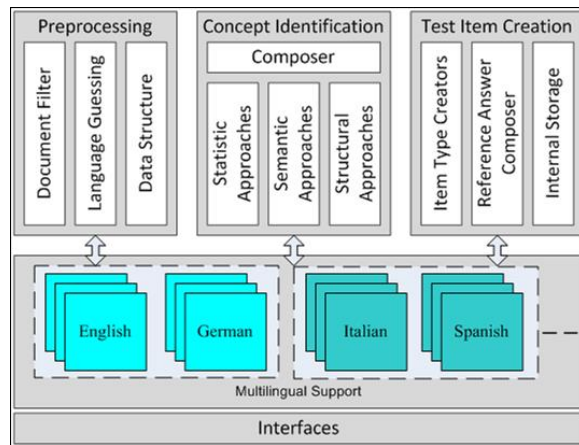


Fig. 2: Conceptual Architecture of EAQC.

provides the functionality to create test items on the basis of concepts which are important for the learner. Therefore, EAQC has been enhanced to enable manual selection of concepts based on learner interest (see Fig. 1). This means that learners are able with the enhanced version of EAQC to manually annotate learning content with the important concepts they aim to learn. Manually selected concepts (based on the annotated learning material) are then used by the EAQC algorithm to automatically create test items. For instance, during the learning process, the learner might highlight the most important concepts from a text and EAQC will use these concepts to automatically create relevant test items. With the resulting test items the learner may then self-assess her/his knowledge or comprehension about the text. Using the EAQC in such individual way has motivated our research interest in whether the EAQC provides such functionality in a sufficient way.

This paper presents an enhanced automated question creator (EAQC) that is able to create test items fully automatically from a textual learning content or semi-automatically based on manually provided concepts. Results show that the overall quality of the test items semi-automatically created by EAQC was comparable to manually created items by humans. The rest of this paper is organized as follows: Section II explains EAQC architecture and EAQC main modules, Section III presents two studies that have been conducted to evaluate EAQC improved functionality, Section IV discusses the studies' main results and findings, whereas, Section V concludes this research.

## II. ENHANCED AUTOMATED QUESTION CREATOR (EAQC)

EAQC utilizes an automated process to create different types of test items out of textual learning content, more precisely EAQC automatically creates single choice, multiple-choice, completion exercises and open ended test items (questions). EAQC is capable of processing textual learning content stored in various file formats, extracting most important related concepts, creating different types of test items and reference answers that adhere to the IMS Question & Test Interoperability (QTI) Specification<sup>1</sup>. As depicted in Fig. 2

EAQC architecture supports multilingual test item creation, currently English and German are supported, whereas a flexible extension to other languages such as Italian and Spanish is possible.

From a conceptual point of view, EAQC consists of three main modules [24] (a) the Pre-processing module, (b) the Concept Identification module and (c) the Test Item Creation module. The modules are explained as follows:

- 1) During the pre-processing module, EAQC detects the input material language (i.e. English or German), performs text cleaning and processing such as special characters and stop words removal, tokenization, and then converts the given learning content into an internal XML format for further processing. Several file formats and online resources are supported by EAQC (e.g. Microsoft Word, Open Document, PDF files, and HTML web content).
- 2) During the concept identification phase, a syntactic analysis based on part-of-speech tagging (POS) is applied using state-of-the-art natural language processing procedures for the identified language (i.e. English or German). This is followed by statistical analysis of term weighting based on terms co-occurrences. Furthermore, semantic word analysis using WordNet [28] is also performed. Results of the processed information further annotate the XML representation of the learning content. The GATE [29] text processing framework was used as part of this phase. The input text is analyzed into tokens, and then POS tagging, named-entity-recognition (NER), text chunking (noun-based), and co-reference resolution analysis are performed for each token. In the statistical analysis, the importance of nouns in the content is estimated and accordingly relevant candidates of word phrases (concepts) are extracted from the learning content. These candidates of word phrases are semantically analyzed using WordNet and prioritized by a relevance number based on the learning content. A configurable threshold value finally defines the concepts in descending order to be used in the third task - i.e. the assessment item creation. Special version of Wordnet is called GermaNet [30] was used for the semantic word analysis on the German text.
- 3) During test item creation phase, EAQC determines the most appropriate sentence out of the learning content for each of the previously extracted concepts and adds two neighboring sentences to the respective sentence in order to provide sufficient context information. In addition, EAQC computes the distractors for multiple-choice items and the antonyms for (incorrect) single-choice items by also making use of the previously outlined statistic, syntactic and semantic analyses of the concepts. Finally, EAQC creates question items applying a template approach and reference answers for the open-ended items and transforms all resulting items into IMS QTI Specification compliant format.

<sup>1</sup>[https://www.imsglobal.org/question/qti2p1/imsqti\\_oviewv2p1.html](https://www.imsglobal.org/question/qti2p1/imsqti_oviewv2p1.html)

### III. EVALUATION AND EXPERIMENTATION

Before testing whether EAQC is able to create test items based on manually extracted concepts, the concepts extracted by learners difference and relevance to the concepts extracted by EAQC were investigated. Therefore, in Study 1, data from [24] was reanalyzed. The concepts that participants had extracted from a text and the concepts extracted by EAQC using the same text were compared. Findings from Study 1 reanalysis were then used in Study 2 in order to investigate whether EAQC is able to create valid test items – not only on the basis of automatically extracted concepts but also on the basis of concepts extracted by learners. To this end, EAQC was fed with the manually created concepts from the participants of Study 1 and the resulting (semi-automated created) test items were evaluated by comparing their quality with fully automatically created ones.

#### A. Study 1

The aim of Study 1 was to investigate whether the concepts extracted by the EAQC from a learning textual material match the concepts extracted by learners from the very same text (Did learners extract similar concepts as the EAQC in terms of content and number?). To this end, data from a study presented in [24] was reanalyzed. The aim of that study was to evaluate the quality of the concepts and test items that were extracted by the EAQC from a text about 'natural-language processing' (NLP). Before participants evaluated the items, their familiarity with the learning topic – their ability to adequately assess the concepts and test items – was assured. Therefore, participants were required to read the text in order to extract - from their viewpoint - the most relevant concepts, then they had to create different types of test items from the text (see below). That is, participants followed similar phases as the EAQC does during test item creation. Comparing the concepts extracted by the participants with the concepts extracted by the EAQC from the same material should allow us to evaluate the level of agreement between the two approaches.

For a better understanding, although the analysis for the concept extraction will be provide afterwards only as outlined above, the full methodology of the study conducted in [24] is presented as follows. Such detailed description of the methodology of Study 1 is also important since the method of the subsequent study (Study 2) is fairly similar to the method of Study 1.

1) *Participants*: 29 participants (4 female) took part in this study. They were 25.4 years on average (SD = 3.3), ranging from 22 to 39 years. All of them were technical students. 93.1% of them already had a bachelor degree. The experiment took place within a regular course Information Search and Retrieval at Graz University of Technology. Because of the restricted number of computer-work places, the participants were divided in two groups (14 and 15 participants each, respectively). The two groups were tested separately on two consecutive days. All participants gave informed consent.

2) *Stimuli and Procedure*: In advance of the study, EAQC was used to create test items from a learning content about 'Natural-Language Processing' (NLP). The text was taken (with slight changes) from Wikipedia<sup>2</sup> and consisted of ap-

<sup>2</sup>NLP:[http://en.wikipedia.org/wiki/Natural\\_language\\_processing](http://en.wikipedia.org/wiki/Natural_language_processing)

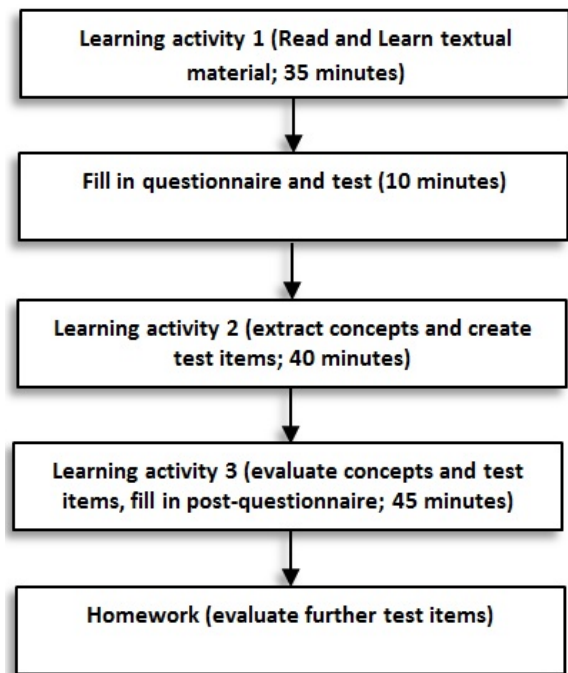


Fig. 3: Flowchart explains the procedure of the experiment.

proximately 2,600 words by that date. EAQC extracted 49 main concepts from this text. Example concepts were, for instance, "natural language processing"; "modern NLP algorithms", and "the Georgetown experiment". The concepts were automatically ranked with regard to their relevance by means of statistical and semantic analyses of the content (i.e., the first concept extracted by EAQC was statistically the most relevant one etc.; see Table 2, left column for the 10 most relevant concepts extracted by EAQC). For each of these concepts, EAQC then created four types of test items (one open-ended item, one completion exercise, one single-choice item and one multiple-choice item, respectively) and the respective answers. In the following, the resulting questions and answers for the concept "modern NLP algorithms" are presented (please note that for the sake of brevity the answer for open-ended items is not provided):

- 1) **Open-ended item:**  
Example: What do you know about modern NLP algorithms in the context of Natural language processing?  
Region of answer: (...)
- 2) **Completion exercise:**  
Example: (...) are grounded in machine learning, especially statistical machine learning.  
Answer: "modern NLP algorithms".
- 3) **Single-choice item:**  
Example: Old style NLP algorithms are grounded in machine learning, especially statistical machine learning.  
Answer: False.
- 4) **Multiple-choice item:**  
Example: (...)are grounded in machine learning, es-

pecially statistical machine learning.

- a) Answer 1: meta-rule NLP algorithm
- b) Answer 2: algorithmic program NLP algorithms
- c) Answer 3: modern NLP algorithms
- d) Answer 4: heuristic NLP algorithms

Together, the EAQC created 196 test items in total (49 concepts  $\times$  4 test-item types). However, not all of the test items were evaluated during the study. In order to reduce the time and effort for the participants, they evaluated only 80 test items which were based on the 20 most relevant extracted concepts (i.e., 20 concepts  $\times$  4 test-item type = 80 test items). In addition, 24 test-items (six per test-item type) were provided. These items had been extracted by one of the authors in advance for checking purpose. Hence, each participant evaluated 104 test items in total (24 during the study session and 80 as a homework assignment). Furthermore, participants evaluated the relevance of the 49 automatically extracted concepts (and seven manually, by one of the authors, extracted concepts). The learning content (i.e., the text), the questionnaires for the evaluation tasks, and also the instructions were presented in English as web-based content. To collect the data, participants were asked to fill in five online questionnaires that were delivered successively at specific times during the study (see below). For creating these questionnaires LimeSurvey<sup>3</sup> was used. Participants were not aware about the immediate purpose of the study (i.e., evaluation of automatically extracted concepts and test items) but were told that they have to take part in several learning activities during the session. Crucially, they were not informed that most of the concepts and test items for evaluation were based on EAQC. Hence, participants did not know the source (i.e., if they were created by EAQC or human). Results from the tests delivered during the experiment were part of the final grading of the course. However, the participation in the experiment was not a prerequisite to complete the course successfully.

The study session consisted of several phases (see Fig. 3). At the beginning of the session, an outline about the session was presented by one of the authors. Then the learning content was presented online (web-based) to each participant. Participants were then asked to learn the prepared text about 'Natural-Language Processing' (Learning activity 1). After that, they had to fill in a questionnaire that was created with LimeSurvey in advance and sent to them via email. In this questionnaire participants were asked not only for specific demographic data such as age, previous education etc. but also to briefly summarize the text they had learned before. Then, after a short break, and crucially for the present paper, participants were required to extract relevant concepts from the text and to create eight test items (two of each test-item type as described before) using their extracted concepts (Learning activity 2). This learning activity aimed at ensuring the participants' familiarity with the text. The activity lasted about 40 minutes. The text was accessible to the students during this task and they were allowed to take notes if necessary. Participants inserted their answers into a new questionnaire that was delivered via email. After they had completed Learning activity 2, participants had to attend a second test in which

they had to answer eight multiple choice items regarding the learning content.

After a further break, Learning activity 3 started. In this learning activity, participants were asked to evaluate the relevance of the 56 concepts (49 automatically extracted concepts and seven manually for control) using a 5-point Likert scale (5 = very relevant; 1 = not relevant at all). Participants were also asked to evaluate the 24 test items ((5 = very good; 1 = very bad) regarding their pertinence (i.e., relevance of the test item with respect to the major themes of the text), level (i.e., Is the test items trivial or does it expresses a significant meaning?), and terminology (appropriateness of the chosen words; see [26]). In addition, if an answer was provided, they had to evaluate the quality of the answer (i.e., Is the provided answer relevant?) and, in case of multiple-choice items, the quality of the distractors was also evaluated. The order of the concepts and test items to be evaluated was randomized. Finally, after completion of Learning activity 3, participants had to fill in a post-questionnaire in which they were asked to answer more general questions about the task (e.g. general attitudes regarding the different test-item types). In total, the experiment lasted approximately three hours. In addition, the evaluation of further test items (16 EAQC-based and four manually created test items per test-item type, resulting in 80 test items in total) was set for homework.

**3) Results and Findings:** Results of the evaluation task and test performance are presented in [24], [27]. For the purpose of this paper, only the concepts which the participants extracted during the experimental session (Learning activity 2) were needed.

All in all, the participants extracted 153 different concepts (491 in total) and 17.1 on average (SD = 10.3; ranging from 5 to 41 concepts per participant). Hence, participants typically extracted fewer concepts than the EAQC (49 concepts). Table 1 (middle column) depicts the 10 most frequently extracted concepts by the participants and the 10 statistically most relevant concepts extracted by the EAQC (left column). More than 90% of the students extracted "machine learning" or "natural language processing" whereas about 38% chose "named entity recognition". When two independent raters were asked to rate which of the concepts extracted by EAQC perfectly match a concept extracted by the students they agreed only on 9 (out of 147) perfect matches (e.g., "natural language processing", "machine learning"); in further 57 cases there was some disagreement between the raters because the concepts matched at least partially (e.g., "word/text segmentation" and "evaluation"). However, most of the time (in 87 cases) participants extracted concepts that were not considered by EAQC (e.g., "parsing", "word sense disambiguation").

In sum, the findings of Study 1 suggest that there is some overlap between the concepts created by the EAQC and manually extracted concepts. However, students often also extracted concepts that were not considered by the tool. In addition, they considered fewer concepts as relevant as the tool. Hence, not all concepts extracted by the EAQC might be experienced as important for the individual learner and some perhaps relevant concepts are even missed. It may be therefore useful to allow the individual learner to enter his or her own concepts to the EAQC in order to receive appropriate test items. Therefore the EAQC was improved in order to support such functionality. In

<sup>3</sup><http://www.limesurvey.org/>



TABLE I: Most important concepts extracted by the EAQC (left column) and by the participants in study 1(middle column; percentage of naming in parentheses). For a better comparison, also the concepts extracted by the participants of Study 2 are presented (right column).

Concepts by EAQC	Study 1: Concepts by students	Study 2: Concepts by students
natural language processing	machine learning (96.5%)	natural language processing (87.5%)
modern NLP algorithms	natural language processing (93.1%)	machine learning (75.0%)
languages text segmentation	part-of-speech tagging (72.4%)	artificial intelligence (50.0%)
the first statistical machine translation systems	NLP evaluation (65.5%)	linguistics (37.5%)
linear algebra and optimization theory	parsing (48.3%)	NLP evaluation (37.5%)
computer science and linguistics	word sense disambiguation (48.3%)	NLP tasks (37.5%)
machine learning	statistical NLP (41.4%)	Turing test (25%)
the Georgetown experiment	word segmentation (41.4%)	hand-written rules (25%)
evaluation metrics	topic segmentation and recognition (41.4%)	fully automatic translation (25%)
an evaluation step	Named entity recognition (37.9%)	statistical NLP (25%)

such scenario, user-defined concepts are used instead of the automatically extracted concepts. The remainder of the test-item creation procedure remains the same.

### B. Study 2

To investigate the quality of the test items that are based on manually provided concepts, a second study was conducted. In this second study semi-automated test items (i.e., automatically generated test items that were based on manually extracted concepts) with fully automated test items (i.e., test items that were based on concepts extracted by the EAQC) were compared. Furthermore, a sample of test items for control that were created completely manually was included.

1) *Participants*: Eight participants took part in this study (2 out of them were females). Participants were 33.1 years on average (SD = 6.6), ranging from 25 to 41 years. 87.5% of them were PhD students in computer science and 12.5% were master students at Graz University of Technology. All participants gave informed consent.

2) *Apparatus and Stimuli*: Apparatus and Stimuli were the same as in the first study except for the tool enhancement of processing user-defined concepts for the test item creation process. To create the test items using EAQC, the concepts provided by EAQC and the concepts that had been extracted by the participants during the first study were used. In particular, the 10 – out of 15 – most frequently extracted concepts from the participants in the first study were used to create test items using EAQC. Some of the participants extracted concepts had to be slightly rephrased in order to create test items automatically out of them as there was no fuzzy matching mechanism implemented in that version of EAQC.

In total, 120 test items were presented based on three categories with 40 test items each (10 per each of the four test-item types). The first category included test items fully provided by the EAQC (EAQC-a). These items were the same as in the first study. The second category consisted of 40 items created by the EAQC based on the concepts that had been extracted by the participants in the first study (EAQC-m). Finally, the third category (manual) included a randomly selected sample of the test items that had been created by the participants during the second learning activity in Study 1 (see above). For this latter category only fully elaborated test items were considered (i.e. in which the participants of

Study 1 provided not only the question but also the respective - correct - answer).

3) *Procedure*: The procedure of Study 2 was similar to the first study with the following exceptions. The study was not conducted during a course but participants attended the study as an online session. In addition, participants had to evaluate 120 test items in total. There was no time restriction for the tasks although participants were given the same guidelines as described in the first study (see Fig. 1). Participants were also asked to complete a further questionnaire in which they had to rate the quality of multiple-choice questions distractors. As this task was of no relevance for the aims of the current study, results of this questionnaire are not reported here.

4) *Results and Findings*: First, the concepts that were extracted by the participants were analyzed and compared with the concepts extracted by the participants in Study 1 and the concepts extracted by EAQC. Participants in Study 2 extracted 53 different concepts (100 in total) and 12.5 on average (SD = 8.7; ranging from 3 to 24 concepts per participant). Table 1 shows the 10 most frequently extracted concepts, by EAQC (which were the same for both studies in column 1), by the participants during the first study (column 2), and by the participants during the second study (column 3). Despite the fact that fewer participants took part in Study 2, the concepts were quite similar to the concepts extracted by the participants Study 1. That is, in both studies, the most important concepts were "natural language processing" and "machine learning". Also "NLP evaluation" and "statistical NLP" were mentioned by both groups. However, there were also concepts that were only considered by the participants of one study (e.g., "parsing" and "history of NLP" only in the first study vs. "artificial intelligence" and "linguistics" only in the second study). This could be - at least to some extent - attributed to differences in prior knowledge between the two groups (note that most of the participants in the second study were PhD students) and is further evidence that learners differ slightly in their individual views of which concepts are relevant.

For each test-item type, the quality of 10 EAQC-created items based on automatically extracted concepts (EAQC-a), 10 EAQC-created items based on manually extracted concepts (EAQC-m), and 10 fully manually created test items were analyzed. The evaluation criteria were the same as in the first study (see [26]). That is, participants evaluated the

TABLE II: Mean ratings for test items: automatically created based on automatically extracted concepts (EAQC-a), automatically created based on manually extracted concepts (EAQC-m), and manually created test items (Manually) for each test-item type with regard to the evaluation criteria. Standard deviations are presented in parentheses.

Question Type	Approach	Pertinence	Terminology	Level	Answer	Distractors
Completion Exercises	EAQC-a	3.7 (0.7)	3.6 (0.7)	3.4 (0.7)	3.7 (0.7)	-
	EAQC-m	3.6 (0.7)	3.6 (0.7)	3.4 (0.6)	3.7 (0.8)	-
	Manually	3.7 (0.9)	3.7 (0.7)	3.6 (0.8)	3.7 (0.7)	-
Single-Choice Items	EAQC-a	3.7 (0.7)	3.8 (0.8)	3.7 (0.7)	-	-
	EAQC-m	3.5 (1.0)	3.5 (0.5)	3.4 (0.8)	-	-
	Manually	3.3 (0.8)	3.3 (0.8)	3.2 (0.8)	-	-
Open-Ended Items	EAQC-a	3.9 (0.7)	3.6 (0.7)	3.9 (0.5)	3.6 (0.6)	-
	EAQC-m	3.9 (0.7)	3.8 (0.6)	3.9 (0.6)	3.6 (0.6)	-
	Manually	4.2 (0.7)	4.2 (0.7)	4.0 (0.6)	3.8 (0.6)	-
Multiple-Choice Items	EAQC-a	3.6 (0.6)	3.6 (0.6)	3.3 (0.7)	3.5 (0.5)	3.1 (0.8)
	EAQC-m	3.6 (0.8)	3.5 (0.7)	3.1 (0.7)	3.4 (0.7)	2.9 (0.8)
	Manually	3.8 (0.7)	3.9 (0.6)	3.6 (0.8)	3.9 (0.8)	3.9 (0.7)

pertinence, the terminology and the level of the test items as well as, when appropriate, the relevance and quality of the answers and distractors. Table 2 shows mean ratings (1 = very bad; 5 = very good) for each test-item type with regard to the evaluation criteria. In order to investigate possible quality differences between the three item's sources (EAQC-a; EAQC-m, and manual, respectively) regarding the evaluation criteria repeated-measures ANOVAs separately for each test-item type were computed. In case the assumption of sphericity was violated, the Greenhouse-Geisser correction was used to correct for degrees of freedom.

For completion exercises, the ANOVA showed no effect of item source,  $F < 1$ , but a main effect of evaluation criteria,  $F(3, 21) = 4.66$ ,  $p < 0.05$ ,  $\eta_p^2 = 0.40$ , and no interaction,  $F < 1$ . This finding suggests that the quality of both categories of automatically created completions exercises was comparable to the manually created items.

For single-choice items, a main effect of item source was found,  $F(2, 14) = 7.78$ ,  $p < .01$ ,  $\eta_p^2 = 0.53$ , but no effect of evaluation criteria,  $F(2, 14) = 2.94$ ,  $p = .09$ , and no interaction,  $F < 1$  were found. Post-hoc analysis (Bonferroni corrected) showed that manually created items were evaluated even worse compared to automatically created items based on automatically extracted concept ( $p < .05$ ). No difference in quality was found between the manually created items and EAQC-m items and between the two types of automatically created items.

For open-ended items, a main effect of item source was found,  $F(2, 14) = 5.88$ ,  $p < .05$ ,  $\eta_p^2 = 0.46$ , and a main effect of evaluation criteria as well,  $F(1.33, 9.33) = 4.74$ ,  $p < .05$ ,  $\eta_p^2 = 0.40$ . There was no interaction,  $F(1.69, 11.82) = 1.01$ ,  $p = .38$ . Post-hoc analysis showed that manually created items were evaluated better compared to EAQC-m items. There was also a tendency ( $p = .09$ ) that manually generated test items were evaluated better than EAQC-a items. There was again no difference between EAQC-m and EAQC-a items.

Finally, for multiple-choice items, a main effect of item source was found,  $F(2, 14) = 7.58$ ,  $p < .01$ ,  $\eta_p^2 = 0.52$ , and also a main effect of evaluation criteria,  $F(4, 28) = 6.13$ ,  $p < .01$ ,  $\eta_p^2 = 0.47$ . The interaction was also significant,  $F(8, 56) = 2.70$ ,  $p < .05$ ,  $\eta_p^2 = 0.28$ . Post-hoc analysis showed that manually created items were again evaluated better as

compared to EAQC-m items ( $p < .05$ ). There was no such difference between manually and EAQC-a items and between EAQC-a and EAQC-m items.

#### IV. DISCUSSION

In order to evaluate the quality of EAQC automatically extracted concepts and test items two studies were conducted. In the first study, whether and to what extent learners might extract the same concepts as EAQC out of a learning content were tested. Results showed that there was an overlap between the manually and automatically extracted concepts. In addition, participants extracted on average fewer concepts than EAQC. Together, this suggested that not all concepts provided by EAQC might be relevant for learners. Therefore, the second study tested whether it is possible to personalize EAQC in such a way that the tool has the functionality to create valid test items out of concepts that were manually selected and inserted to the tool. To this end, EAQC was fed with manually extracted concepts and compared the quality of the resulting test items with test items that were either fully automatically or fully manually created.

Results showed that, in general, the quality of both the semi- and fully automatically created test items was comparable to manually created ones. In particular, for completion exercises, both types of automatically created items did not differ from the manually created test items with regard to various quality criteria such as pertinence, terminology, and level. For single choice items, the automatically created test items were evaluated even slightly better than the manually created items. However, there were two test-item types in which manually created items outperformed their (fully and semi-) automated created counterparts: open-ended items and multiple-choice items. For multiple-choice items, the results might be caused by the relatively low quality of the distractors created by EAQC. In the same time for multiple-choice items the participants' ratings for manually and (both fully and semi-) categories differed only slightly with regard to all other quality criteria (pertinence, terminology, level of the test items and the relevance of the answer). The average ratings for the quality of EAQC-based distractors were approximately one point lower than for the manually created distractors (see Table 2; note that, because of a possible lack of statistical power, the respective post-test did not yield a significant difference). Anecdotal

reports of the students also suggest that the distractors were sometimes too easy. They stated that learners might be able to guess the correct answers by simply excluding the obviously inappropriate alternatives.

Previous research has shown that creating appropriate and valid distractor items is in fact demanding (e.g. [31], [32]). The creation of appropriate distractors can be difficult even if the distractors are created manually. For instance, DiBattista and Kurzawa [33] showed that many distractors created for various classroom tests were flawed and needed revision. Creating meaningful distractors using question-creation tools is even more challenging (e.g., [17]). In general, good distractors should be as semantically close to the correct answer as possible [34]. Currently, the EAQC uses antonyms and related terms of the respective concept in order to compute distractors. This approach might, however, lead to the creation of distractors that are clearly odd when presented in a specific context (e.g., correct answer: "speech tagging"; example distractor computed by the EAQC: "tongue-lashing tagging"). Thus, further experimentation is necessary to improve the quality of distractors automatic creation as part of multiple-choice questions.

With regard to the open-ended items, the cause of the worse quality of both types of automatically created test items as compared to manually created test items is not that clear as for the multiple-choice items. One reason for the poor evaluation of the quality of these items might be the "uniform" terminology of the main phrase that scaffolds open-ended items. That is, for automatically created test items, the standard phrasing used was "What do you know about [concept] in the context of (or subjected to) NLP?". In contrast, the structure of the manually created items was more diversified (e.g., "Explain.....", "What is [the difference between] ...?", "Describe....."). Perhaps this made the test items more interesting and less "artificial" for the participants. However, this does not mean that participants were not able to understand the meaning of the questions created by EAQC. The results suggest that the automatically created test items were nevertheless syntactically valid. Hence, even if the open-ended automatically created questions may be less quality than human created ones, they still can be used for self-assessment without difficulty.

In Study 2, test items that were either based on fully automatically extracted concepts or on manually provided ones have been evaluated. Note that, as the content of the automatically and manually extracted concepts sometimes overlapped –i.e. they use similar terminology (see Table 1) – this had created typically similar test items (e.g., EAQC created the same test items for the concept 'natural-language processing' which had been extracted by both the EAQC and the participants in Study 1). It can be assumed that such overlap between test items of different item categories (semi automatically vs. fully automatically) led to similar ratings of the respective test items, and hence, the lack of finding differences between the quality of these two test item categories. From a technological and methodological point of view, such finding reflects the validity of EAQC architecture, when the same concepts had led to the creation of the same test items. However, from a pedagogical point of view, a possible enhancement of the tool might be that other alternative items are computed from the same concept. Such variation in the test items might prevent having the same

test item being presented again and again when learners self-assess their learning progress more than once. Moreover, it might prevent learners of becoming bored or even being less challenged during learning.

The number of concepts extracted by the participants in both studies varied from about five concepts per person to more than 40 concepts per person. Hence, there were large differences between individual learners with regard to how many concepts they considered as relevant. Noticeably, the EAQC extracted far more concepts (49 concepts) from the same learning content than the average learner did. Thus, there might be a mismatch between the "statistical" relevance of a concept and its relevance from the learner's individual point of view. In Study 1, only 9 (i.e., 6.1%) out of the 147 concepts that were extracted by the students in total were perfectly consistent with the concepts extracted by the EAQC. The same analysis for Study 2 revealed a perfect match for only 2 (i.e., 3.8%) out of 53 concepts. This is insofar critical as the test items that result on these concepts might then also be often "worthless" for a learner.

The EAQC already provides the opportunity to reject less important concepts after the concept-extraction phase (i.e., before starting the actual test-item creation, see Fig. 1). Therefore, learners may read through all concepts the EAQC has provided and have the EAQC create test items only from the most appropriate concepts. This functionality (which was purposefully not used in the present studies) may already help to improve the quality of the automatically created test items. In addition, it is also possible to reduce the number of concepts that the EAQC is supposed to extract from the learning content before the concept-extraction phase (i.e., the learner can determine the number of concepts that should be extracted).

From a pedagogical point of view, it might not always make sense to reduce the number of concepts in advance, because such "less relevant" concepts might sometimes cover aspects that the learner had missed so far during learning. Rather, it might be suitable to have EAQC first extract the concepts based on a statistical analysis from which the learner can then select the most appropriate ones. This process of reflecting on the concepts extracted by EAQC might help learners to deepen their understanding of the text. In a final step, learners should then be allowed to include concepts that are still missing from their viewpoint. Based on these three steps, the resulting test items created by EAQC should then cover the main topics of the learning content which in turn are assumed to support learners efficiently. This paper shows that EAQC is partially able to support this functionality outlined above. The tool allows selecting specific concepts from the extracted ones and it was able to create test items from 10 out of 15 manually inserted concepts. However, some of the concepts had to be slightly rephrased in order to receive valid test items. Such rephrasing is of course quite time-consuming. Thus, further improvements for fuzzy term matching and synonym and hypernym relations are considered in this regard.

## V. CONCLUSION

In the context of self-regulated learning learners often lack adequate feedback of whether they have covered all aspects of



a topic or whether the important points they have extracted are consistent with the points proposed by colleagues or instructors [10]. As a consequence, they often have problems to create appropriate test items for self-assessment.

In this paper an enhanced automated question creator (EAQC) is presented. EAQC is able to either create test items fully automatically from a learning content or on the basis of manually provided concepts. Two studies evaluating EAQC improved functionality were discussed. The results showed that the overall quality of the test items semi-automatically created by EAQC was comparable to manually created items. These findings are a further step in developing a tool that is able to effectively support learners during self-regulated learning and self-assessment activities.

## REFERENCES

- [1] T. Janssen, "Instruction in self-questioning as a literary reading strategy: An exploration of empirical research," *L1-Educational Studies in Language and Literature*, vol. 2, no. 2, pp. 95–120, 2002.
- [2] A. King, "Effects of self-questioning training on college students' comprehension of lectures," *Contemporary Educational Psychology*, vol. 14, no. 4, pp. 366–381, 1989.
- [3] M. Scardamalia and C. Bereiter, "Text-based and knowledge based questioning by children," *Cognition and instruction*, vol. 9, no. 3, pp. 177–199, 1992.
- [4] T. Papinczak, R. Peterson, A. S. Babri, K. Ward, V. Kippers, and D. Wilkinson, "Using student-generated questions for student-centred assessment," *Assessment & Evaluation in Higher Education*, vol. 37, no. 4, pp. 439–452, 2012.
- [5] D. Buehl, *Classroom strategies for interactive learning*. International Reading Assoc., 2013.
- [6] N. Silveira, "Towards a framework for question generation," in *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, 2008.
- [7] A. C. Graesser, S. E. Gordon, and L. E. Brainerd, "Quest: A model of question answering," *Computers & Mathematics with Applications*, vol. 23, no. 6, pp. 733–745, 1992.
- [8] B. S. Bloom and M. D. Engelhart, *Taxonomy of Educational Objectives: The Classification of Educational Goals: By a Committee of College and University Examiners: Handbook 1*. David McKay, 1969.
- [9] L. W. Anderson, D. R. Krathwohl, and B. S. Bloom, *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Allyn & Bacon, 2001.
- [10] C. K. E. Cheng, "The role of self-regulated learning in enhancing learning performance," 2011.
- [11] J. M. Bugg and M. A. McDaniel, "Selective benefits of question self-generation and answering for remembering expository text," *Journal of educational psychology*, vol. 104, no. 4, p. 922, 2012.
- [12] M. Dostál and K. Jezek, "Automatic keyphrase extraction based on nlp and statistical methods," in *DATESO*, 2011, pp. 140–145.
- [13] Y. Lui, R. Brent, and A. Calinescu, "Extracting significant phrases from text," in *Advanced Information Networking and Applications Workshops, 2007, AINAW '07. 21st International Conference on*, vol. 1, May 2007, pp. 361–366.
- [14] M. Agarwal, R. Shah, and P. Mannem, "Automatic question generation using discourse cues," in *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, ser. IUNLPBEA '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1–9. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2043132.2043133>
- [15] D. L. Lindberg, "Automatic question generation from text for self-directed learning," Ph.D. dissertation, Applied Sciences: School of Computing Science, 2013.
- [16] P. Piwek and K. E. Boyer, "Varieties of question generation: introduction to this special issue," *Dialogue and Discourse*, vol. 3, pp. 1–9, 2012.
- [17] T. Goto, T. Kojiri, T. Watanabe, T. Iwata, and T. Yamada, "Automatic generation system of multiple-choice cloze questions and its evaluation," *Knowledge Management & E-Learning: An International Journal (KM&EL)*, vol. 2, no. 3, pp. 210–224, 2010.
- [18] L. Vanderwende, "The importance of being important: Question generation," in *Proceedings of the 1st Workshop on the Question Generation Shared Task Evaluation Challenge*, Arlington, VA, 2008.
- [19] O. Sitthisak, L. Gilbert, and H. Davis, "An evaluation of pedagogical informed parameterized questions for self-assessment," *Learning, Media and Technology*, vol. 33, no. 3, pp. 235–248, 2008.
- [20] J. Villalon and R. A. Calvo, "Concept extraction from student essays, towards concept map mining," in *2009 Ninth IEEE International Conference on Advanced Learning Technologies*. IEEE, 2009, pp. 221–225.
- [21] W. J., "Extraction of relevant semantic data from natural language texts in the view of automatic question generation," 2010.
- [22] N. Liu and C. C. Yang, "Keyphrase extraction for labeling a website topic hierarchy," in *Proceedings of the 11th International Conference on Electronic Commerce*. ACM, 2009, pp. 81–88.
- [23] E. Hovy, Z. Kozareva, and E. Riloff, "Toward completeness in concept extraction and classification," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. Association for Computational Linguistics, 2009, pp. 948–957.
- [24] C. Gütl, K. Lankmayr, J. Weinhofer, and M. Hofler, "Enhanced automatic question creator-eaqc: Concept, development and evaluation of an automatic test item creation tool to foster modern e-education," *Electronic Journal of e-Learning*, vol. 9, no. 1, pp. 23–38, 2011.
- [25] M. Al-Smadi and C. Guehl, "Supporting self-regulated learners with formative assessments using automatically created qti-questions," in *Global Engineering Education Conference (EDUCON), 2011 IEEE*, April 2011, pp. 288–294.
- [26] S. Cannella, E. Ciancimino, and M. López Campos, "Mixed e-assessment: An application of the studentgenerated question technique," in *Education Engineering (EDUCON), 2010 IEEE*, April 2010, pp. 769–773.
- [27] M. Höfler, M. AL-Smadi, and C. Gütl, "Investigating content quality of automatically and manually generated questions to support self-directed learning," in *CAA 2011 International Conference, University of Southampton*, 2011.
- [28] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [29] H. Cunningham, "Gate, a general architecture for text engineering," *Computers and the Humanities*, vol. 36, no. 2, pp. 223–254, 2002.
- [30] B. Hamp, H. Feldweg *et al.*, "Germanet-a lexical-semantic net for german," in *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Citeseer, 1997, pp. 9–15.
- [31] T. M. Haladyna, *Developing and validating multiple-choice test items*. Routledge, 2012.
- [32] J. R. Moser, C. Gütl, and W. Liu, "Refined distractor generation with lsa and stylometry for automated multiple choice question generation," in *AI 2012: Advances in Artificial Intelligence*. Springer, 2012, pp. 95–106.
- [33] D. DiBattista and L. Kurzawa, "Examination of the quality of multiple-choice items on classroom tests," *Canadian Journal for the Scholarship of Teaching and Learning*, vol. 2, no. 2, p. 4, 2011.
- [34] R. Mitkov, H. LE AN, and N. Karamanis, "A computer-aided environment for generating multiple-choice test items," *Natural Language Engineering*, vol. 12, no. 02, pp. 177–194, 2006.

# Characterizing End-to-End Delay Performance of Randomized TCP Using an Analytical Model

Mohammad Shorfuzzaman and Mehedi Masud  
Computer Science Department  
Taif University, Taif, Saudi Arabia

Md. Mahfuzur Rahman  
Dept. of Computer Science and Information Technology  
Hajee Mohammad Danesh Science and Technology University  
Dhaka, Bangladesh

**Abstract**—TCP (Transmission Control Protocol) is the main transport protocol used in high speed network. In the OSI Model, TCP exists in the Transport Layer and it serves as a connection-oriented protocol which performs handshaking to create a connection. In addition, TCP provides end-to-end reliability. There are different standard variants of TCP (e.g. TCP Reno, TCP NewReno etc.) which implement mechanisms to dynamically control the size of congestion window but they do not have any control on the sending time of successive packets. TCP pacing introduces the concept of controlling the packet sending time at TCP sources to reduce packet loss in a bursty traffic network. Randomized TCP is a new TCP pacing scheme which has shown better performance (considering throughput, fairness) over other TCP variants in bursty networks. The end-to-end delay of Randomized TCP is a very important performance measure which has not yet been addressed. In the current high speed networks, it is increasingly important to have mechanisms that keep end-to-end delay within an acceptable range. In this paper, we present the performance evaluation of end-to-end delay of Randomized TCP. To this end, we have used an analytical and a simulation model to characterize the end-to-end delay performance of Randomized TCP.

**Keywords**—Randomized TCP, end to end delay, congestion window, TCP pacing, propagation delay, Markov chain.

## I. INTRODUCTION

Transmission Control Protocol (TCP) plays a very important role in data transmission over network. TCP offers a connection oriented delivery service to the end user applications and it provides reliable data flows between two processes running on two end systems. TCP implements data retransmission mechanism for the packets that are lost during transmission and thus ensures guaranteed data transfer between the sender and the receiver. TCP is intelligent enough to understand the loss of packets that are already sent by implementing *timeout* timer at the sender and by using duplicate Acknowledgements messages. TCP implements slow start and congestion avoidance phases to handle the congestion of the network [1]. However, it cannot fully avoid the packet losses which can greatly degrade the throughput quality of the transmission system. This problem becomes very severe when the network is bursty. The network may have bursty traffic very often due to buffer overflow and the limited buffer size of intermediate routers. Researchers found that if the router buffer size can be made equal to the product of link bandwidth and average Round Trip Time (RTT) of flows passing through the router, then the packet loss ratio will be reduced and this will improve the end-to-end delay performance [2]. But this large amount

of memory implementation will not be feasible and costly as well. Research is also going on to modify some aspects of TCP algorithm addressing this problem. TCP pacing is proposed in which successive data packets are transmitted with some time intervals in between which can avoid sending packets in bursts. TCP pacing uses the last Round Trip Time (RTT) to adjust not only its next congestion window size (how much) but also the time (when) of sending. The congestion window is the amount of data the sender is allowed to send to the network at a time. Randomized TCP improves the paced TCP to achieve better result. In [3], Chandrayana et al. has proposed Randomized TCP which randomizes the packet sending time at sender. Randomized TCP is very similar to paced TCP but in paced TCP packet sending time at TCP sources are equally spaced for all flows where in Randomized TCP the packet sending times are scheduled at different intervals or randomly for the TCP flows. Randomized TCP solves the phase effect and biasness problems which are still present in TCP pacing. Randomized TCP performs better than paced TCP in improving network throughput, fairness, timeouts and losses. But all these proposals do not address the end-to-end delay performance achieved by Randomized TCP [3]. To this end, this paper evaluates the performance of Randomized TCP in end-to-end delay. It is highly desired that the Randomized TCP shows better end-to-end delay performance than paced TCP. The evaluation shows both network and application layer performance of Randomized TCP.

The remainder of the paper is organized as follows. We have discussed the necessary background in Section 2. Related work relevant to the proposed research and the research problem are presented in Section 3. Simulation and analytical modeling are presented in Section 4 and Section 5 respectively. Finally, Section 6 concludes the paper with some future work.

## II. BACKGROUND

### A. Congestion Window

Each sending endpoint of a TCP connection possesses a buffer for storing data which will be transmitted over the network. On the other hand, receiver side buffering helps the application to read the data only when it is ready. This also lets network transfers take place while applications are busy with other processing, improving overall performance. TCP uses “flow control” to avoid overflowing the receiver side buffer. In this case, TCP sends a fixed number of packets at a time and the size is usually known ‘congestion window’. This congestion window (cwnd) contains the amount

of data that may be transmitted from the sender buffer. TCP uses “congestion control” mechanism to dynamically control congestion window size. In TCP, the window size depends on the congestion of the network. If there is any packet loss due to congestion in the network then the window size becomes very small otherwise it increases for each successful packet transmission. This congestion control window have a great impact in the performance of network. In TCP pacing and Randomized TCP, the sender delays the sending of packets in addition to the changing of congestion window when it identifies any congestion in the network.

## B. TCP Algorithms

TCP is a very dynamic and reliable congestion control protocol. It uses acknowledgments and the TCP acknowledgments created by the destination are returned to the source. TCP acknowledgments help the sender to know whether the packets are well received or not. In TCP transmission, lost packets are interpreted as congestion signals. There are a number of standard variants of the TCP protocol, such as Tahoe, Reno, New Reno, Sack and Vegas. One of the main differences between these TCP versions lies in their methods of recovering from packet loss due to network congestion. In general, they differ in the sender side algorithms. FAST TCP is a new high speed TCP protocol which uses the experienced queuing delay of the packets to adjust the congestion window size. In our work, we have used TCP NewReno as a general variant of TCP to compare the end-to-end delay performance of Randomized TCP. A detailed account of some of these protocols is given below.

While TCP Reno produces less bursty traffic than TCP Tahoe, it is much less robust towards phase effects. The latter refers to unpredictability in performance resulting from very small differences in the relative timings of packet arrivals for different connections sharing a link. Both versions of TCP appear to have significant drawbacks as a means of providing data services over multimedia networks, because random loss resulting from fluctuations in real-time traffic can lead to significant throughput deterioration in the high bandwidth-delay product. The performance is degraded when the product of the loss probability and the square of the bandwidth delay product is large.

For high bandwidth-delay products, TCP is hideously inequitable towards connections with higher propagation delays: for multiple connections sharing a bottleneck link, the throughput of a connection is inversely proportional to (a power of ) its propagation delay. It is worth expounding that random loss causes performance decline in TCP because it does not permit the TCP window to reach high enough levels to allow superior link utilization. On the other hand, when the TCP window is already big and is causing congestion, random early drops of packets when the link buffer gets too filled can in fact improve performance and lessen phase effects [4].

Relatively earlier simulation studies of TCP-tahoe include [5], [6], [7]. Simulations for the simple multi-hop network considered in [6] showed the oscillations in window sizes and the inequality of TCP towards connections traversing a larger number of hops. In [5], the authors consider a number of TCP connections sharing a bottleneck link. There is no

queuing of acknowledgements, and sources are assumed to have data to send at all times. As stated earlier, [7] considers the effect of two-way traffic.

The unfairness of TCP-tahoe against connections with large round-trip delays and against connections traversing a big number of congested gateways has also been demonstrated in other current studies of TCP-tahoe. The heuristic analysis shows that, for multiple connections sharing a bottleneck link, the throughput of a connection is inversely proportional to its round-trip time. Oscillatory behavior and inequality towards connections with superior propagation delays have also been noticed in an earlier analytical study of feedback-based congestion control which uses a continuous-time approximation to the dynamic behavior of a rate-based scheme.

Another adaptive window flow control scheme is proposed in [8], [9]. The proposed window adaptation mechanism operates in a high bandwidth-delay product, and is based on asymptotics derived from a queuing model of the network. It has the disadvantage of requiring more central synchronization than TCP: the adaptation algorithm for every link must be acquainted with the distinctiveness of the bottleneck link for that connection and the relative propagation delays of the other connections sharing that link. However, the adaptive mechanism itself is much smoother than the drastic window size changes in TCP, so that a decentralized adaptive scheme based on a similarly smooth mechanism may defeat the weaknesses in TCP while not requiring the type of acquaintance understood in [8], [9].

## C. TCP pacing and Randomized TCP

TCP’s congestion control mechanisms can not avoid bursty traffic flows on high-speed networks. Bustry Traffic produces higher queuing delays, more packet losses, lower throughput. TCP pacing is aimed at reducing the burstiness of TCP traffic and the impact of limited buffers in routers. In TCP pacing, the packet loss is less and the competing flows face less queuing delays. In addition to setting TCP’s congestion window which tells about *how much to send*, TCP pacing also fixes the time *when to send* the packets. In TCP pacing, the sender sends successive data packets with some equally spaced time intervals,  $\Delta = \frac{RTT}{cwnd}$ . In this case, Round Trip Time (RTT) and cwnd are the key components to adjust the sending time of packets. Though TCP pacing reduces packet losses and queuing delays, this scheme is not efficient in addressing TCP’s “phase effect” problem and it also has biasness against long flows. When there are a number of competing flows, phase effect causes a specific section of competing flows to experience recurrent drops. Randomized TCP is found efficient to address this problem and it also reduces the biasness against long flows. In Randomized TCP, the sender sends successive data packets with randomly spaced time intervals,  $\Delta = \frac{RTT(1+x)}{cwnd}$ , where x follows a uniform distribution. Randomized TCP also shows more fairness over Paced TCP when multiplexed with other standard TCP variants (e.g. TCP New Reno) [3]. All the researches include traffic performance such as fairness, throughput of Randomized TCP. But more application-layer performance like end-to-end delay has not been addressed. We have evaluated the end-to-end delay performance of Randomized TCP in this paper. In a recent work, the authors in [10] propose a transport layer solution to

TCP called TCPRand. The idea is to have a randomization of TCP payload size, which breaks synchronized packet arrivals between flows from different input ports. The simulation results show that TCPRand ensures the improvement of TCP fairness with negligible overheads in all test cases. Prakah et al. [11] investigate the TCP outcast problem in data center applications with many-to-one traffic pattern scenario. The research results found out that with excessive traffic flows, drop-tail queueing may drop a series of consecutive packets at each input port having a port blackout. Wu et al. [12] focus on incast congestion control for TCP in data center networks using a similar traffic pattern scenario.

#### D. End-to-end delay

End to end delay can be defined as the time taken by a packet to reach the destination after it has started from the source. If the network is not busy then there will not be any queueing delay. But if the network has bursty traffic then the queueing delay also affects the end-to-end delay performance. Processing time of the routers, transmission delay also affect the end-to-end delay of packets. We have used an analytical model to characterize the end-to-end delay performance of Randomized TCP and also evaluate the results with simulation.

#### E. NS-2 Simulation Environment

NS-2 is a discrete event simulator which has mainly targeted at networking research. NS-2 provides substantial simulation support for wired and wireless (local and satellite) networks. NS-2 provides tools to create any desired networking environment and also to import any networking aspects (e.g. TCP, routing, and multicast etc.) for data transmission in that environment. NS (version-2) is developed under the VINT project as a joint effort by UC Berkeley, USC/ISI, LBL, and Xerox PARC. NS-2 is an object oriented simulator and was written in C++ with OTcl as a front-end. The simulator handles two class hierarchy: the C++ class hierarchy (compiled hierarchy), and the class hierarchy within the OTcl interpreter (interpreted hierarchy) [13]. We have used NS-2 network simulator for experiments.

### III. RELATED WORK

Many applications rely on TCP since TCP has its own congestion control mechanism and does not bother the network for that. In [14], self regulating TCP acknowledgement pacing scheme has been proposed. This work tells about the ACK pacing technique to reduce the data loss due to congestion in TCP. In TCP each packet needs to be acknowledged by the receiver. But in acknowledge pacing scheme, the acknowledges are delayed by the receiver. ACK pacing uses a matrices (network load dynamics) to make the receiver understand about the congestion and allows the receiver acts upon this situation very smartly. Thus, the authors [14] show the implementation issues and better performance results of self-regulating ACK pacing.

In [15], Rezdan et al. tell about TCP Westwood which uses Bandwidth Share Estimate (BSE) along with RTT to set the pacing interval for sender side TCP pacing. TCP Westwood keeps the bottleneck service rate in BSE parameter. There are two major phases of TCP: slow start phase and congestion

avoidance phase. In this work, TCP pacing is only effective in the slow start phase. In [16], Garetto et al. present a way to analyze the bursty TCP traffic in wide area networks and present ways to characterize "TCP pacing". The authors use a simple analytical model to show the traffic (e.g. packet loss) produced by a large number of TCP connections. In [17], Chang et al. find out that paced TCP performs better channel reuse than TCP Reno.

In [18], ElRakabawy et al. introduce gateway adaptive pacing scheme which is mostly effective for the flows from wired sender to wireless receivers. The gateway device implements a pacing queue and it sends the queued data obtained from the wired sender with understanding of the current transmission rate of the wireless network. This approach includes some transport layer functionality to the IP layer in the Internet gateway. In our work, TCP NewReno is used as standard TCP protocol.

Enachescu et al. has found that buffer size of routers can be made very small if some performance of link layer can be sacrificed [19]. FDL (Fiber Delay Line) provides limited buffering capacity. In [20], TCP pacing is measured using RTT/cwnd which shows TCP pacing application performance in the FDL buffers. In this work, the authors find that TCP Pacing reduces packet loss-rate by decreasing the burstness of packet arrival and achieves higher throughput. [21] shows the effects of small buffer in standard TCP as well as TCP pacing. In our evaluation, It is found that TCP pacing is very useful in small buffer routers. In recent work, the authors in [22], [23] have focused on multipath TCP traffic which enables hosts to send data over multiple paths and has use cases on smartphones and datacenters. The research results confirm that multipath TCP operates successfully over the real Internet even with the middleboxes.

Now, we describe the problem definition for our current research. TCP does not bother network to handle the flow control and congestion control rather it is handled by the end systems. But inside the network the congestion mainly occurs due to buffer overflow at routers. When the network is congested, the TCP sender only adjusts the congestion window and as a result it cannot avoid the packet losses. So, it is very important to make the TCP sender to delay in sending packets when the network has bursty traffic. If the TCP sender delays in sending the packets, then it can reduce packet losses. TCP pacing has addressed this issue and Randomized TCP improves the limitations of paced TCP. In this paper, We have evaluated the end-to-end delay performance of Randomized TCP through simulation and analytical modeling.

### IV. SIMULATION MODELING

#### A. Simulation Setup

We used NS-2 network simulator to simulate the network environment. As shown in Figure 1 [24], we considered the topology of Abilene Network (Internet2). In Abilene network, there are 11 nodes [node(0,1,2,...,10)] and 14 links. There is a bottleneck link between node(2) and node(6). The channels between node(2) and node(6) have only 45Mbps bandwidth whereas all other channels possess 155Mbps bandwidth. We used the table I for assigning the propagation delay of each channel. In simulation, we used 'DropTail queues' and 'fixed

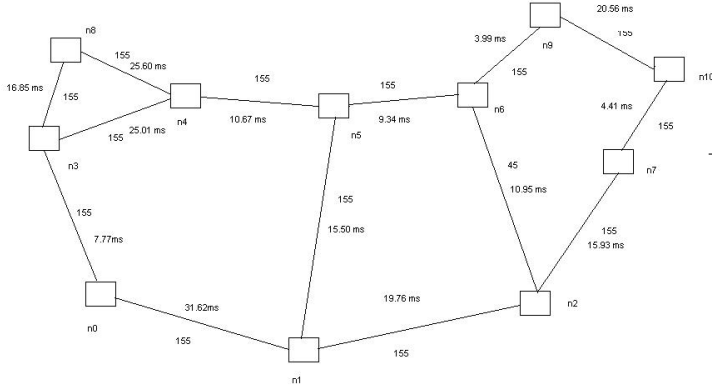


Fig. 1: Network Topology [24]

Source	Destination	Pro. Delay
n(0)	n(1)	31.624ms
n(0)	n(3)	7.772ms
n(1)	n(2)	19.756ms
n(1)	n(5)	15.504ms
n(2)	n(6)	10.950ms
n(2)	n(7)	15.938ms
n(3)	n(4)	25.010ms
n(3)	n(8)	16.852ms
n(4)	n(8)	25.608ms
n(4)	n(5)	10.674ms
n(5)	n(6)	9.340ms
n(6)	n(9)	3.990ms
n(7)	n(10)	4.412ms
n(9)	n(10)	20.464ms

TABLE I: Propagation Delays [24]

routing’. Drop Tail queue drops newly arriving packets until the queue finds enough room to accept incoming traffic. In fixed routing, there is a unique path for each source-destination pair.

In this simulation, nodal delays are neglected since processing delays incurred by packets are comparatively very less than propagation delays in channels. Using sample run, we characterized the traffic for the network to obtain expected utilization of each channel. NS-2 is an open source simulator and we modified the NS-2 source code (existing TCP implementation) to implement Randomized TCP. For this, we modified the ‘delay’ variable in send-much function (in tcp.cc) by introducing uniform randomization. TCP uses the ‘delay’ variable to measure the next packet sending time when the network is bursty. TCP fixes up the delay value as  $\frac{RTT}{cwnd}$  but we introduced a uniform random variable with this to obtain the characteristics of Randomized TCP. So in case of Randomized TCP, the delay value is measured by  $\frac{RTT(1+x)}{cwnd}$ , where x follows a uniform distribution. TCP uses ‘t\_rt\_’ variable for saving RTT values and ‘cwnd\_’ variable saving for current congestion window size. TCP NewReno

Queue Size	Randomized	NewReno
20 pkts	49.97 ms	51.51 ms
35 pkts	46.86 ms	48.47 ms
50 pkts	45.29 ms	46.76 ms
65 pkts	44.02 ms	44.90 ms
5000 pkts	43.49ms	45.79 ms

TABLE II: End-to-end Delays for different queue size

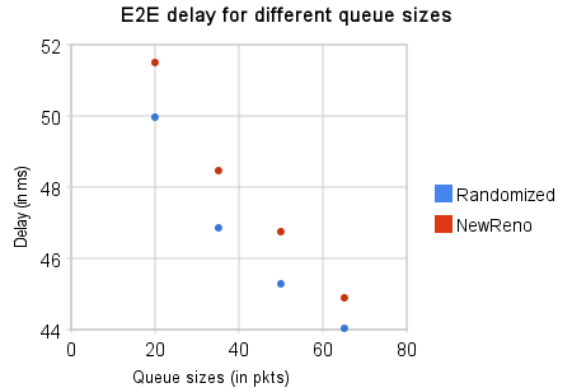


Fig. 2: End-to-end delays for different queue size

is an already developed module in NS-2. We used TCP NewReno to compare the performance of Randomized TCP. In all cases, delay is calculated as the difference between successful reaching time of each packet at destination and the starting time from the source. We have only considered the ‘tcp packets’ for calculating end-to-end delay, and the ‘ack packets’ and ‘connection establishment packets’ are not considered for delay calculation. We have varied the total number of flows in the network to control the burst level of the network.

### B. Experiments and Results

First of all, we measured the end-to-end delays of Randomized TCP and TCP NewReno by varying the queue size of channels. The results shown in Table II tell that Randomized TCP’s performance is peak over TCP NewReno when the queue size is very small. With the increase of queue sizes, the delay difference between Randomized TCP and TCP NewReno becomes narrow or small. In case of infinite buffer, Randomized TCP also shows good performance over TCP NewReno. In all these cases, we fixed up the total flow number as 90 and each packet size as 1000 Bytes.

In the next experiment, we fixed up the queue size as small (size as 35) and have varied the total number of flows to change the amount of burst in network. It is found that, for this small queue size and in high bursty traffic, Randomized TCP is far better than TCP NewReno. But in infinite buffer size (we used queue size as 5000) consideration, the burstiness cannot make significant delay difference between Randomized TCP and TCP NewReno. We depicted the results in Table III and in Table IV.

From the above simulation results, it is found that Randomized TCP is very good in considering end-to-end delay per-

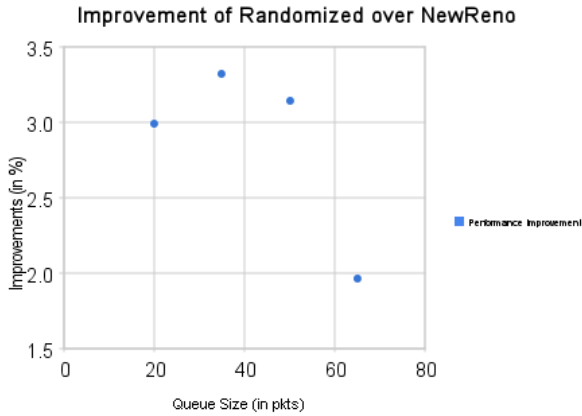


Fig. 3: Performance improvement in Randomized TCP

Total Flow	Randomized	NewReno
45	43.24 ms	43.34 ms
90	46.86 ms	48.47 ms
135	64.34 ms	107.62 ms

TABLE III: E2E Delays for different bursts (Queue size=35: as small size)

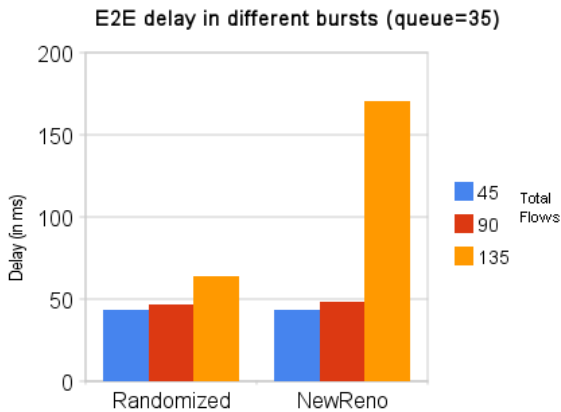


Fig. 4: E2E Delays for different bursts (Queue size=35)

Total Flow	Randomized	NewReno
45	39.78 ms	39.78 ms
90	43.49 ms	45.79 ms
135	57.28 ms	58.38 ms

TABLE IV: E2E Delays for different bursts (Queue size=5000: as infinite size)

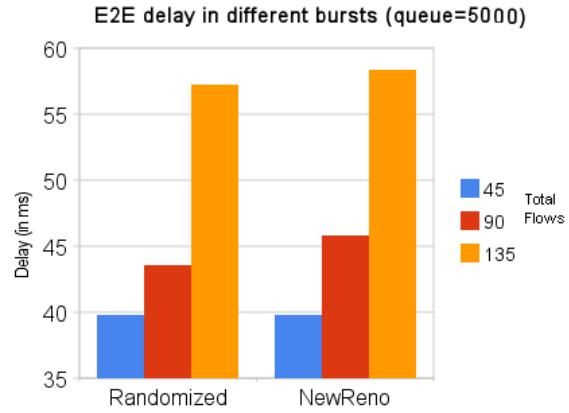


Fig. 5: E2E Delays for different bursts (Queue size=5000)

formance. About 3.25% delay improvements can be achieved for small queue size (Figure 3). Again in bursty network, Randomized TCP also shows better result over TCP NewReno when the channels have small buffer (Figure 4, 5).

## V. ANALYTICAL MODELING

This analytical model is very similar to the queueing network model proposed by Lam et al. [25]. It is assumed that the routers in the network are indexed by  $i = 1, 2, 3, 4, \dots, M$ . Each router works on First Come First Serve (FCFS) basis and works at constant rate  $C_i$  bits per second. There may be more than one flow and the flows are indexed by  $k = 1, 2, 3, \dots, K$ . The first-order Markov Chain with transition probabilities for a particular flow  $k$  is modeled as  $p_{ij}^k$  which is the routing probability to server  $j$  from server  $i$  and  $i, j = 1, 2, 3, 4, \dots, M$ .

It is assumed that the packets from  $k$  flows arrive at source node following poisson distribution with  $\gamma_k$  (packets per second), where  $k = 1, 2, 3, 4, \dots, K$ . The total external arrival rate to the network,

$$\gamma = \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \dots + \gamma_K. \quad (1)$$

For a particular flow  $k$ , if  $\gamma_k$  and  $p_{ij}^k$  are given then at router  $i$  the arrival rate of packets (due to a particular flow  $k$ ),  $\lambda_{ik}$  is identified as

$$\lambda_{ik} = \gamma_k \delta_{ik} + \sum_{j=1}^M \lambda_{jk} p_{ji}^k, \quad (2)$$

where  $\delta_{ik}$  is 1 when  $i$  is source node of flow  $k$ , otherwise  $\delta_{ik}$  remains zero. So, the total arrival rate at router  $i$  (due to all flows),

$$\lambda_i = \sum_{k=1}^K \lambda_{ik}. \quad (3)$$

Again, the traffic intensity at router  $i$  for a particular flow  $k$ ,

Total Flow	Analytical	Simulation
45	33.35 ms	39.78 ms
90	51.86 ms	45.59 ms
135	198.48 ms	57.28 ms

TABLE V: Simulation and Analytical Results

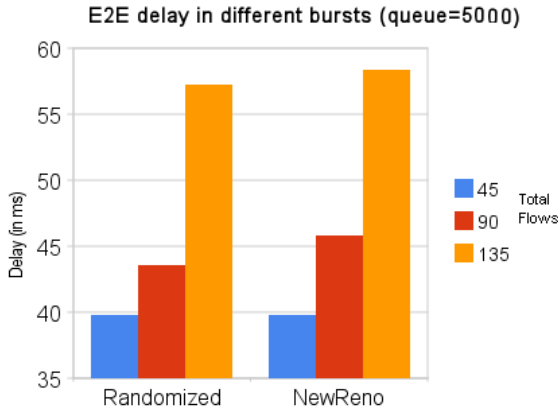


Fig. 6: Analytical vs Simulation Results

$$\rho_{ik} = \frac{\lambda_{ik}}{\mu C_i}, \quad (4)$$

where  $\mu$  is used as mean in poisson distribution. The total traffic intensity at router  $i$  for all flows,

$$\rho_i = \sum_{k=1}^K \rho_{ik}. \quad (5)$$

This  $\rho_i$  can be used to find out the mean number of packets in transit within the network,

$$\gamma T = \sum_{i=1}^M \frac{\rho_i}{1 - \rho_i} = \sum_{i=1}^M \frac{\lambda_i}{\mu C_i - \lambda_i}, \rho_{ik} = \frac{\lambda_{ik}}{\mu C_i}. \quad (6)$$

So, the mean end-to-end delay,

$$T = \frac{1}{\gamma} \sum_{i=1}^M \frac{\lambda_i}{\mu C_i - \lambda_i}. \quad (7)$$

#### A. Evaluation

The analytical model assumes customer arrival to follow a poisson distribution and also assumes the packet size to follow exponential distribution. But TCP as well as Randomized TCP does not follow these distributions in packet sending. So there arose a big difference in analytical and simulation results. We showed both results in Table V and Figure 6

## VI. CONCLUSION

TCP is an important aspect in Computer Networking. Randomized TCP algorithm finds out a way of obtaining useful and automated services in data communication. This can introduce the best usage of the existing resources. The end-to-end delay evaluation in this paper makes Randomized TCP a more attractive and useful Transport layer solution. We compared the end-to-end delays of Randomized TCP and TCP NewReno by varying the queue size of channels. The results show that Randomized TCP's performance is peak over TCP NewReno when the queue size is very small. With the increase of queue sizes, the delay difference between Randomized TCP and TCP NewReno becomes narrow or small. In case of infinite buffer, Randomized TCP also shows good performance over TCP NewReno. Thus, the users can easily enjoy their required best services through Randomized TCP. As a future work, we are planning to work on different queuing algorithms with Randomized TCP congestion control mechanism and specifically implementing a new queue mechanism which may give better performance in Randomized TCP congestion control than the existing queue mechanisms.

## REFERENCES

- [1] J. F. Kurose and K. W. Ross, *Computer Networking: A top down approach featuring the internet*. Addison Wesley, 2002.
- [2] G. Hasegawa, T. Tomioka, K. Tada, and M. Murata, "Simulation studies on router buffer sizing for short-lived and pacing TCP flows," *Computer Communications*, vol. 31, no. 16, pp. 3789 – 3798, 2008. [Online]. Available: <http://www.sciencedirect.com/science/article/B6TYP-4SGKBBD-3/2/cd284735365599e5edd17eac89380d2e>
- [3] K. Chandrayana, S. Ramakrishnan, B. Sikdar, and S. Kalyanaraman, "On randomizing the sending times in TCP and other window based algorithms," *Computer Networks*, vol. 50, no. 3, pp. 422–447, 2006.
- [4] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Transactions of Networking*, vol. 1, no. 4, pp. 397–413, 1993.
- [5] S. Shenker, L. Zhang, and D. D. Clark, "Some observations on the dynamics of a congestion control algorithm," *Computer Communication Review*, pp. 30–39, 1990.
- [6] L. Zhang, "A new architecture for packet switching network protocols," Ph. D. dissertation, M.I.T. Lab. Comput. Sci., Cambridge, MA, 1989.
- [7] L. Zhang, S. Shenker, and D. D. Clark, "Observations on the dynamics of a congestion control algorithm: the effects of two-way traffic," in *Proc. ACM SIGCOMM '91*, 1991, pp. 133–147.
- [8] D. Mitra, "Asymptotically optimal design of congestion control for high speed data networks," *IEEE Trans. Commun.*, vol. 40, no. 2, pp. 301–311, 1992.
- [9] D. Mitra and J. B. Seery, "Dynamic adaptive windows for high speed data networks with multiple paths and propagation delays," *Computer Networks and ISDN Systems*, vol. 25, pp. 663–679, 1993.
- [10] S. Lee, M. Lee, D. Lee, H. Jung, and B. S. Lee, "Tcprand: Randomizing tcp payload size for tcp fairness in data center networks," in *Proc. of IEEE Conference on Computer Communications (INFOCOM)*, 2015, pp. 1697 – 1705.
- [11] P. Prakash, A. Dixit, Y. C. Hu, and R. Kompella, "The tcp outcast problem: Exposing unfairness in data center networks," in *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*. Berkeley, CA, USA: USENIX Association, 2012, pp. 30–30.
- [12] W. Haitao, F. Zhenqian, G. Chuanxiong, and Z. Yongguang, "Ictcp: Incast congestion control for tcp in data center networks," in *Proceedings of the 6th International Conference*, ser. Co-NEXT '10. New York, NY, USA: ACM, 2010, pp. 13:1–13:12.
- [13] F. K and V. K., "Notes and documentaiton, LBNL," <http://www.isi.edu/nsnam/ns/>.
- [14] J. Aweya, M. Ouellette, and D. Y. Montuno, "A self-regulating TCP



- acknowledgement (ACK) pacing scheme,” *International Journal of Network Management*, vol. 12, no. 3, pp. 145–163, 2002.
- [15] A. Razdan, A. Nandan, R. Wang, M. Sanadidi, and M. Gerla, “Enhancing TCP performance in networks with small buffers,” in *Proceedings of the 11th IEEE International Conference on Computer Communications and Networks*. IEEE, 2002, pp. 39–44.
- [16] M. Garetto and D. Towsley, “An efficient technique to analyze the impact of bursty TCP traffic in wide-area networks,” *Performance Evaluation*, vol. 65, no. 2, pp. 181 – 202, 2008. [Online]. Available: <http://www.sciencedirect.com/science/article/B6V13-4NWCGRD-1/2/0fab98f6270f2f0706849a94e42091bd>
- [17] C.-Y. Luo, N. Komuro, K. Takahashi, and T. Tsuboi, “Paced TCP: A dynamic bandwidth probe TCP with pacing in adhoc networks,” in *International Symposium on Personal, Indoor and Mobile Radio Communications PIMRC*. IEEE, 2007, pp. 1–5.
- [18] S. M. ElRakabawy, A. Klemm, and C. Lindemann, “TCP with gateway adaptive pacing for multihop wireless networks with internet connectivity,” *Computer Networks*, vol. 52, no. 1, pp. 180 – 198, 2008. [Online]. Available: <http://www.sciencedirect.com/science/article/B6VRG-4PT29C0-2/2/a5366ee033e4e1a9b831190ad1c0b581>
- [19] M. Enachescu, Y. Ganjali, A. Goel, N. McKeown, and T. Roughgarden, “Part III: routers with very small buffers,” *SIGCOMM Computer Communication Review*, vol. 35, no. 3, pp. 83–90, 2005.
- [20] O. Masafumi, I. Maroto, and T. Tatsuro, “Application of TCP pacing to optical packet network with Fiber Delay Line buffers,” in *Proceedings of the 7th IEEE International Conference on Optical Internet*. IEEE, 2008, pp. 1–2.
- [21] D. Wischik, “Buffer sizing theory for bursty TCP flows,” in *International Zurich Seminar on Communications*. IEEE, 2006, pp. 98–101.
- [22] Q. Coninck, M. Baerts, B. Hesmans, and O. Bonaventure, “Evaluating android applications with multipath tcp,” *Mobicom 2015*, pp. 230–232, 2015.
- [23] B. Hesmans, H. Tran-Viet, R. Sadre, and O. Bonaventure, “A first look at real multipath tcp traffic,” *Lecture Notes in Computer Science*, vol. 9053, pp. 233–246, 2015.
- [24] T. Lakshman and U. Madhow, “The performance of tcp/ip for networks with high bandwidth-delay products and random loss,” 1997. [Online]. Available: [citeseer.ist.psu.edu/lakshman96performance.html](http://citeseer.ist.psu.edu/lakshman96performance.html)
- [25] U. of Texas TR-069, “Queueing network model,” <http://www.cs.utexas.edu/ftp/pub/techreports/tr81-167.pdf>.

# Improving Vertical Handoffs Using Mobility Prediction

Mahmoud Al-Ayyoub, Ghaith Husari and Wail Mardini  
Jordan University of Science and Technology  
Irbid, Jordan

**Abstract**—The recent advances in wireless communications require integration of multiple network technologies in order to satisfy the increasing demand of mobile users. Mobility in such a heterogeneous environment entails that users keep moving between the coverage regions of different networks, which means that a non-trivial vertical handoff scheme is required in order to maintain a seamless transition from one network technology to another. A good vertical handoff scheme must provide the users with the best possible connection while keeping connection dropping probability to the minimum. In this paper, we propose a handoff scheme which employs the Markov model to predict the users' future locations in order to make better handoff decisions with reduced connection dropping probability and number of unnecessary handoffs. Through simulation, the proposed scheme is compared with the SINR-based scheme, which was shown to outperform other vertical handoff schemes. The experiments show that the proposed scheme achieves significant improvements over the SINR-based scheme that can reach 51% in terms of the number of failed handoffs and 44% in terms of the number of handoffs.

**Keywords**—Heterogeneous wireless networks, Vertical handoff, Markov model, Artificial intelligence, Mobility management.

## I. INTRODUCTION

Even before the first commercial adoption of cellular network in North America and Western Europe in the late 70s and early 80s, such networks received a great deal of attention from researchers in both academia and industry [14]. The recent wide spread of smart phones and tablets allowed people to connect to the Internet easily. Coupled with the increase in the demand for online multimedia material and Voice over IP (VoIP) applications, the need to maintain a "high quality" connection to the Internet anywhere and anytime has become a very important issue. This is achieved through the use of multiple wireless access technologies. One example are the Wireless Local Area Networks (WLANs), which are widely deployed because they are cheap to deploy and provide high bandwidth. However, contrary to Wide Code Division Multiple Access (WCDMA) networks, WLANs have very low service mobility due to its shorter transmission range with limited area of coverage. This fosters the necessity for collaboration between heterogeneous wireless networks such as WLANs, Wide Code Division Multiple Access (WCDMA), and other network variations. Users, or Mobile Terminals (MTs), have different service needs regarding the connection bandwidth, monetary cost, security, etc. Also, MTs move from one location to another that might be outside the area of coverage of the preferred wireless network currently serving the MTs. So, a MT should get the best possible connection that suits its needs

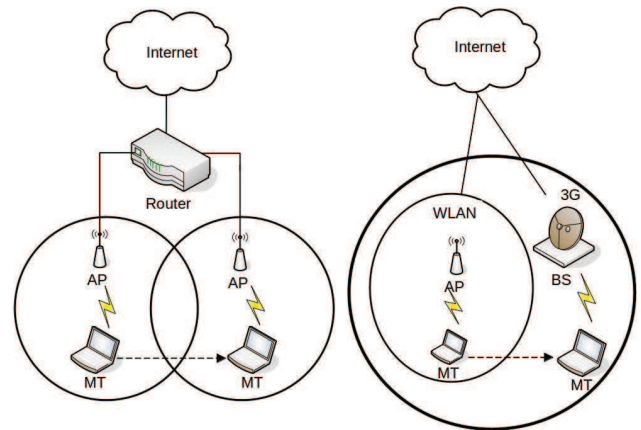


Fig. 1: The two types of handoff: horizontal (left side) and vertical (right side).

and maintains the Always Best connected (ABC) concept as it affects the Quality of Service (QoS) of the connection [8].

Different access technology devices have different transmission power levels and thus different areas of coverage. When a MT starts moving away from its connection-providing device, another device that covers the new location of the MT is needed. The process of moving a MT's communication session from an access technology device to another in order to ensure the continuity of the connection without affecting the ongoing session is called *handoff* [18], [21]. Two types of handoff are commonly discussed in the literature: (i) horizontal handoff, which occurs when the MT moves from one access point to another within the same wireless technology, and (ii) vertical handoff, which occurs when the MT moves from one access point to another access point that belongs to different wireless technology. Figure 1 shows an example for each of these two types. It is noteworthy to mention that the focus of this work is on the more challenging case where the MT has a list of handoff candidates and it must choose the best one wisely in order to maintain its communication session while trying to achieve higher QoS.

As a MT starts moving away from the service providing device, the Received Signal Strength (RSS) gets weaker and the service quality starts to degrade. So, at a certain point, the MT must change its association and connect to a new access point that can provide better service (e.g., more bandwidth, stronger signal strength, etc.) without the loss of the ongoing call or connection. In the literature [1], [4], there have discussions of other factors governing the handoff decision including the

Signal to Interference plus Noise Ratio (SINR), power, delay, available bandwidth and QoS guarantees, security and financial issues. The focus of this work is on improving the reliability of the connection by reducing the probability of disconnections, which occur when MTs try to perform a handoff to an access technology device with no available channels. Disruptions of this type are generally viewed as annoying and unacceptable so much that the users are willing to accept lowered bandwidth if it means improved reliability [23]. Another objective of the proposed scheme is to minimize the number of unnecessary handoff leading to better QoS guarantees [23].

Despite the strong intuition that the geographical nature of the handoff problem plays a significant role in the success of any handoff scheme, most of the current works simply ignore it. This work is based on the idea that the knowledge of the future positions of the MT can aid in making better handoff decisions. One of the benefits gained from this is choosing a handoff destination that is usable for a longer period of time, and thus, reducing the number of unnecessary handoffs. Another benefit is realized when considering the fact that the intersections of movement trajectories of the many MTs are bound to create heavily loaded regions. Taking the future locations of the MT can allow the handoff scheme to avoid choosing the handoff candidates with such “central” locations, which helps in balancing the load in the network as well as decreasing the disconnection probability. The experiments discussed in Section IV support these intuitive arguments as to why the proposed scheme outperforms other schemes such as the SINR-based scheme.

The rest of the paper is organized as follows. The following section surveys the recent works in the literature on vertical handoff schemes while Section III describes the system model used in this work, the assumptions made and the proposed scheme. The performance of the proposed scheme is evaluated in Section IV and the paper is concluded in Section V.

## II. RELATED WORKS

Due to its importance to the increasingly spreading heterogeneous wireless networks, the vertical handoff problem has been studied extensively. Below, we review some of the most important schemes proposed in the literature. We categorize the covered schemes based on their general approach.

We start our coverage with one of the basic techniques which focusing on Received Signal Strength (RSS) as the main factor in the handoff decision. An example of such techniques is the adaptive lifetime-based vertical handoff (ALIVE-HO) scheme proposed by Zahran et al. [25], [24]. While considering delay, authentication, and service initiation, the ALIVE-HO scheme used the RSS to estimate how long a MT's needs can be served by a WLAN which helped in the handoff decision. The objectives were to reduce the number of dropped connections and increase the connection duration by delaying the handoff. According to simulation results shown by the authors, this led to reducing unnecessary handoffs while increasing the average throughput. Moreover, the proposed scheme is able to adapt to the application requirements and user mobility by using the lifetime metric. In a similar work, Yan et al. [19], [20] used RSS measurements to estimate how long a MT will stay in a WLAN cell (i.e., the traveling distance

within the cell). These estimations were exploited in a vertical handoff scheme with reduced number of unnecessary handoffs and improved overall network utilization. The experiments conducted showed that their algorithm reduced the probability of handoff failures as well as unnecessary handoffs provided that the estimated traveling distance is smaller than a certain threshold. Finally, Mohanty et al. [12] proposed a vertical handoff scheme that uses MTs speed and handoff signaling delay to improve the handoff process WLANs and 3G cellular networks.

A closely related criterion to RSS is the signal to interference plus noise ratio (SINR). The vertical handoff scheme of Ayyappan et al. [3] used SINR as the main criterion in the handoff decision. Using Shannon's capacity theorem, the SINR values of each network were used to compute the throughput of each network. The network with the best QoS was selected. The experiments conducted by the authors showed that the SINR-based scheme outperforms other the RSS-based schemes in terms of the throughput and the number of dropped connections. Yang et al. [22] proposed a multi-dimensional adaptive SINR based vertical handoff scheme (MASVH) scheme, which incorporated many aspects into the handoff decision including the SINR, the required user bandwidth, the traffic cost and the network utilization. The conducted experiments showed that MASVH improved the network throughput and decreased both the probability of failed handoffs and the cost of traffic.

In the last scheme of the previous paragraphs, the authors employed (among many factors) the available bandwidth as a factor in the handoff decision. This is not the only work with this basic idea. A QoS-based vertical handoff scheme between WLANs and Wireless Wide Area Networks (WWANs) was proposed by Ayyappan and Kumar [2] which takes into account the available bandwidth as well as the QoS requirements of the user. Another similar scheme for vertical handoff between WLANs and WCDMA networks was proposed by Yang et al. [23]. They key idea in this scheme was to use the SINR values to estimate the achievable bandwidths at the candidate handoff networks in order to make QoS-aware handoff decisions.

As mentioned in Section I, despite the strong intuition that the geographical nature of the handoff problem plays a significant role in the success of any handoff scheme, most of the current works simply ignore it. However, there are a few works (such as [10], [26], [4]) that took geographical information into account while performing the vertical handoff decision. In [10], the locations of the handoff candidates were used to repeatedly compute the distances between them and the MT's position. The handoff decision is triggered when the MT determines that it is moving away from its current access technology device (which can be easily determined if the MT maintains the history of its recent few positions) and the chosen candidate is the closest one to the MT's current position provided that it is closer than current access technology device. Another location-based scheme was proposed by Zhang et al. [26], where dynamic programming is used to utilize the MT's current location and mobility to improve the handoff decision.

Since the proposed handoff scheme utilizes prediction methods to improve the handoff decision, it is important to cover works following a similar approach. Becvar [5] proposed

a new scheme to perform handover based on prediction that uses the history of handover between every two adjacent base stations. With repetitive handovers from a certain base station to an adjacent one, future handovers of mobile terminals that arrive to the base station could be predicted based on the frequency of handovers history between pairs of base stations. The proposed method needs time to adapt and calculate handover probability from a base station to another based on handover history. The paper states that the model needs 2000 to 3000 seconds, which corresponds to 4000 to 6000 handovers [5]. The model must know its environment which consists of neighboring base stations. So, whenever a handover occurs from a base station to another, it affects the probability of handover between that pair of base stations in a positive manner and other base stations handoff probability in a negative manner (if one option probability is increased, other available options probability is decreased). The simulation results show that the prediction hit rate is high when very repetitive handover between a pair of base stations. However, when there are many neighbors and the handover history is distributed on all neighboring base stations, the prediction of handover drops. The prediction hit rate in all scenarios varies from 47% - 20% since it decreases with the increase of neighboring base stations. Finally, Chi et al. [7] suggested two criteria, the wrong decision probability (WDP) and the handover probability (HP), for evaluating handoff schemes and provided mathematical modeling for both measures.

A completely different fuzzy logic approach for vertical handoff between WLANs and Universal Mobile Telecommunications Systems (UMTS) was followed by Xia et al. [17]. The authors suggested using RSS and available bandwidth in addition to predicting the future RSS using differential prediction algorithms to trigger the handoff process. The handoff candidate is selected by the performance evaluation results of a Fuzzy logic based Normalized Quantitative Decision (FNQD).

In a more recent work exploiting fuzzy logic, Boussen et al. [6] proposed a new context aware vertical handover decision algorithm to select the best network in terms of QoS and energy efficiency. They use a fuzzy logic system to initiate handoffs based on context. However, the fuzzy logic system they use is not as accurate as artificial intelligence prediction algorithms (e.g., Markov Model).

Omheni et al. [13] proposed a new approach for handover decision making in heterogeneous wireless network to select the best network based on the application. Simulation showed that the proposed algorithm guarantees QoS requirements and reduce the blocking probability of handoff requests, which also maximizes bandwidth offered from the selected network. However, the approach required assistance from the network to provide context information in case of mobility.

Finally, Wang [15] provided performance evaluation framework for network selection strategies based on models constructed by using stochastic process algebra. The proposed framework captures the throughput rate and mobility features of the nodes in 3G-WLAN interworking networks. Evaluation is performed in terms of throughput, number of handovers, and network blocking rate. They showed that performance is very sensitive to the traffic pattern of the mobile node which affects the number of handoffs that in turn affects the throughput rate and number of blocked handoff requests (i.e., QoS). In

our work, we use the same evaluation parameters to show the quality of our proposed algorithm.

### III. SYSTEM MODEL AND PROPOSED SOLUTION

Before going into the details of the proposed vertical handoff scheme, a quick coverage of the system model and assumptions made is necessary. The following subsection discusses the signal propagation model whereas Section III-B discusses the considered mobility models.

#### A. Signal Propagation Model

This work considers heterogenous networks consisting of two different types of access technology: WLANs and WCDMA cellular networks. As with other wireless communication technology, the main point is the Received Signal Strength (RSS). In the following paragraphs, we discuss the equations used to compute SINR and RSS for both network types under consideration [23], [24], [3], [4].

We start with WCDMA cellular networks. For a WCDMA Base Station (BS)  $j$  and MT  $i$ , the RSS (in dBm) is computed using the following equation.

$$RSS_C = P_j + G_j - PL_{ij} - A_j, \quad (1)$$

where  $P_j$  is the transmission power of  $j$  (in dBm),  $G_j$  is the transmitted antenna gain (in dB),  $PL_{ij}$  is the total path loss (in dB) as defined in Equation 2 and  $A_j$  is the connector and cable loss (in dB). To compute the total path loss for MT  $i$  and BS  $j$ , we use the following equation.

$$PL_{ij} = 135.41 + 12.49 \log(f_j) - 4.99 \log(h_j) + (46.84 - 2.34 \log(h_j)) \log(d_{ij}), \quad (2)$$

where  $f_j$  is the frequency (in MHz),  $d_{ij}$  is the distance (in kilometers) between  $i$  and  $j$  and  $h_j$  is the effective antenna height (in meters).

As for WLANs, the RSS computation is as follows. For a WLAN Access Point (AP)  $j$  and MT  $i$ , the RSS (in dBm) is computed using the following equation.

$$RSS_W = P_j - PL_{ij}, \quad (3)$$

where the total path loss is computed using the following equation.

$$PL_{ij} = L + 10n \log(d_{ij}) + S \quad (4)$$

where  $L$  is a constant power loss,  $n$  is the path loss exponent (assumed to be between 2 and 4), and  $S$  is a zero-mean normal random variable representing the effect of fading.

#### B. Mobility Model

In this work we use two mobility models. The pathway mobility model and a proposed variation of that model which aims to help us better understand the scheme.

Mobile nodes do not move in a completely random way, nor in a complete deterministic way as well. Instead, their movements include both random and regular components [16]. So a memoryless mobility model, that is completely random such as the random waypoint, and a deterministic mobility model (with no randomness as it completely defines the velocity and movement direction for each node in advance and

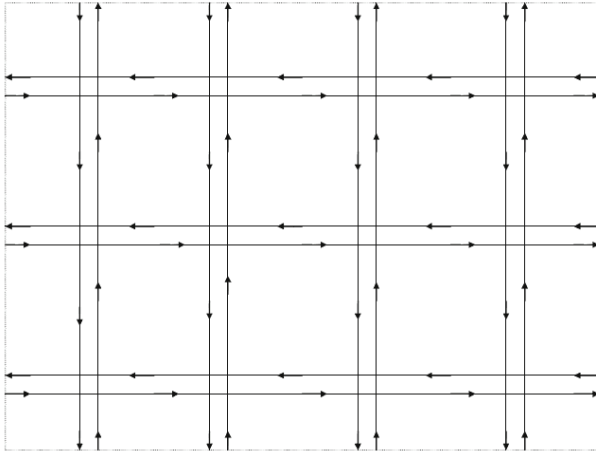


Fig. 2: The Pathway mobility model.

thus every node's movement is certain and known in advance) both are not suitable to describe users' movements inside a city. Accordingly, we describe the movements of MTs using pathway mobility model.

The pathway mobility (PW) model [11] is simple and contains both regular and random components. In this model, the MTs are randomly placed on the edges of the terrain's grid (see Figure 2). Each node randomly chooses a location as a destination point and starts moving to it through the shortest path along the edges of the grid with constant speed chosen uniformly and randomly from a velocity range  $[v_{min}, \dots, v_{max}]$ . When the destination point is reached, the node stops for a randomly selected pause time. After the pause time is over, the node chooses another destination point and repeats the whole process until the simulation time is over. In case there are more than one shortest path, one of them is randomly selected and the mobile node will go to destination through this path.

The pathway model with straight route preference (PWSRP) is similar to the pathway way model. However, the first one picks the shortest route/way to the destination that has the least direction changes (left or right turns). In other words, this model tries to keep going straight as long as possible and avoids making left or right turns unless necessary.

### C. Mobility Prediction-Based Scheme for Vertical Handoff (MPVH)

In this section, the Mobility Prediction-Based Scheme for Vertical Handoff (MPVH) is discussed. MPVH involves two decisions: when should a MT perform a handoff (handoff trigger) and how to do it (i.e., to which BS/AP the MT should connect). Following one of the standard techniques in the literature, a handoff is triggered solely based on the RSS. As a MT  $i$  is moving away from the BS/AP  $j$  to which it is connected, the RSS of  $j$  at  $i$  starts to decrease. At a certain point, the RSS will be too low for the communication to be successful. This will trigger the handoff. Now, the second decision related to which BS/AP the MT should handoff, is where the contribution of MPVH lies. MPVH starts by compiling a list of candidate BSs/APs with a RSS at MT  $i$

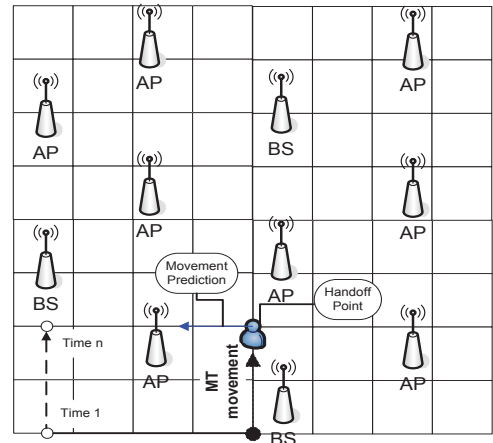


Fig. 3: The handoff process in MPVH.

above the threshold for a successful communication. MPVH then consults the prediction component (which is based on a Markov model) to compute the most probable future location of  $i$  and chooses the BS/AP that is closest to this location. A similar prediction method was used in [9] to improve channel switching in cognitive radio networks. See Figure 3 for an example.

## IV. SIMULATION RESULTS

In this section, simulation experiments were conducted to evaluate the performance of MPVH and compare it with one of the most widely-used vertical handoff schemes. The scheme chosen for comparison is the SINR-based scheme due to its low disconnection ratio in comparison with other schemes as mentioned in Section II.

The two performance metrics used in this comparison are as follows.

- Number of failed handoffs: when a mobile user requests to handoff to an access technology device, the request will be denied if there is no free channels at the access technology device which causes disconnection of the user. These disconnections are very annoying to mobile users. As a matter of fact, users prefer networks with higher reliability (i.e., have lower disconnection probability) to networks with higher bandwidth [23].
- Number of handoffs: Each performed handoff has an associated handoff delay (the time between the last packet received from the old access device and the first packet received from the new access device). The more handoffs performed the more handoff delay there is. This negatively affects the overall QoS [23].

The distribution of access technology devices across the network domain (terrain) directly affects the results of handoff schemes. Previous works [23], [22], [3] have distributed access technology devices across a network domain (terrain) carefully to achieve the best performance of their scheme. We use that topology to compare the MPVH scheme with

the SINR-based scheme, in addition to another topology of uniformly distributed devices. The network consists of 7 Base Stations (BSs) and 12 Access Points (APs). The terrain size is  $5000 \times 5000$  m. In our experiments, the number of MTs varies between 200 and 600. Each MT is placed on a random location and connects to the access technology device of the highest SINR value at the beginning. We use the same network configuration values shown in Table 1 of [4], which were also used in the SINR-based scheme of [23], [3].

The objective of the first experiment is to compare the performance of the two handoff schemes under consideration, MPVH and SINR-based, as the number of MTs increases. The results of this experiment (depicted in Figures 4 and 5) show a clear advantage in favor of MPVH over the SINR-based scheme in both performance metrics.

Considering the number of failed handoffs first, Figure 4(a) shows an average improvement of 35% for the pathway model (PW) and Figure 4(b) shows an average improvement of 43% for the pathway model with straight route preference (PWSRP). Such improvements can be justified as follows. The SINR-based scheme favors the access device with the best SINR value. Such devices may happen to be in “central” locations in the network that are close to the trajectories of movement for many mobile users and hence they usually have high SINR. By favoring such devices, SINR-based scheme is overloading them and possibly leading to more disconnections. On the other hand, MPVH will not favor these devices creating a more load-balanced distribution of work among access devices and reducing the disconnection probability.

As for the number of handoffs, the plots in Figure 5 show that MPVH outperforms the SINR-based scheme by an average of 33% for the PW model and 40% for the PWSRP model (as shown in Figures 5(a) and 5(b)) The interpretation of such results lies in the fact that, under the PWSRP model, MTs avoid changing their directions, which work in favor of MPVH since it uses a movement pattern capturing algorithm for the selection of the best handoff candidate.

To provide more insights into why MPVH outperforms the SINR-based scheme, consider the scenario depicted in Figure 6, where a MT is moving away from its servicing base station ( $BS_1$ ) until it reaches a handoff point (the red circle). At this point, the SINR-based scheme will handoff to the BS that has the best SINR value, which is  $BS_3$ . As the MT changes its directions and moves towards its destination, it reaches another handoff point (the green circle). Then, the SINR-based scheme will be forced to perform another handoff to the BS with the best SINR value at this point, which is  $BS_2$ . On the other hand, MPVH’s mobility prediction component will give it a clear advantage in such scenarios. When the MT reaches the first handoff point (the red circle), MPVH will predict its future location and choose the closest base station to it as the handoff destination, which is  $BS_2$ . Now, as the MT continues its movement, the RSS value of its current servicing BS,  $BS_2$ , will not drop under the threshold and a second handoff will not take place.

In the second experiment, we consider a more realistic (and more challenging) case of random network topology. In the previous experiment, the considered network topology consisted of fixed BS/AP locations as depicted in Figure 3.

Note that the BSs are placed on a triangular grid and the APs are placed in the middle of the overlap regions of the coverage areas of the BSs. Such “perfect” placement gives an advantage to the SINR-based scheme as it allows smaller overlapping regions which reduces the number of candidate handoff destinations. To study this issue, a uniform distribution of the BSs/APs is used. Figures 7(a) and 7(b) show that under such distribution, the average improvements of MPVH over the SINR-based jump to about 44% for the PW model and 51% for the PWSRP model, in terms of the number of failed handoffs. As for the number of handoffs, Figures 8(a) and 8(b) show that under such distribution, the average improvements jump to 37% for the PW model and 44% for the PWSRP model.

## V. CONCLUSION AND FUTURE WORK

In this work, Markov model was used to predict the future location of the user and the vertical handoffs the user will need based on that prediction. We show through simulations and experiments that predicting users’ future locations provides a better tool for handoff than SINR-based scheme. This is evident by the decrease in service disconnection probability (which can reach 51% in certain cases) and the number of unnecessary handoffs (which can reach 44% in certain cases).

There are many future directions of this work. One of them is to use a higher order Markov model to improve the predictions since such model allows the exploitation of longer history of mobile nodes to predict further steps in the future. This is expected to improve the handoff decisions and therefore improve the overall system performance. In order to deal with the added complexity of such models, a compression mechanism will be necessary due the exponential nature of the Markov expansion which will burden the limited memory of a mobile node. Studying the tradeoff between the gain due to having a more accurate prediction model and the cost of maintaining such a system is an interesting future direction of this work.

## REFERENCES

- [1] K. Ayyappan and P. Dananjayan. RSS measurement for vertical handoff in heterogeneous network. *Journal of Theoretical and Applied Information Technology*, 4(10):989–994, 2008.
- [2] K. Ayyappan and R. Kumar. QoS based vertical handoff scheme for heterogeneous wireless networks. *Proceedings of the International Journal of Research and Reviews in Computer Science (IJRRCS'10)*, 1(1):1–6, 2010.
- [3] K. Ayyappan, K. Narasimman, and P. Dananjayan. SINR based vertical handoff scheme for QoS in heterogeneous wireless networks. In *Future Computer and Communication, 2009. ICFCC 2009. International Conference on*, pages 117–121. IEEE, 2009.
- [4] Abdelnasser Banihani, Mahmoud Al-Ayyoub, and Ismail Ababneh. A direction-based vertical handoff scheme. *Journal of Emerging Technologies in Web Intelligence*, 5(3), 2013.
- [5] Z. Becvar. Efficiency of handover prediction based on handover history. *Journal of Convergence Information Technology*, 4(4):41–47, 2009.
- [6] Sinda Boussen, Nabil Tabbane, Sami Tabbane, and Francine Krief. A context aware vertical handover decision approach based on fuzzy logic. In *Communications and Networking (ComNet), 2014 International Conference on*, pages 1–5. IEEE, 2014.
- [7] C. Chi, X. Cai, R. Hao, and F. Liu. Modeling and analysis of handover algorithms. In *Global Telecommunications Conference, 2007. GLOBECOM'07. IEEE*, pages 4473–4477. IEEE, 2007.

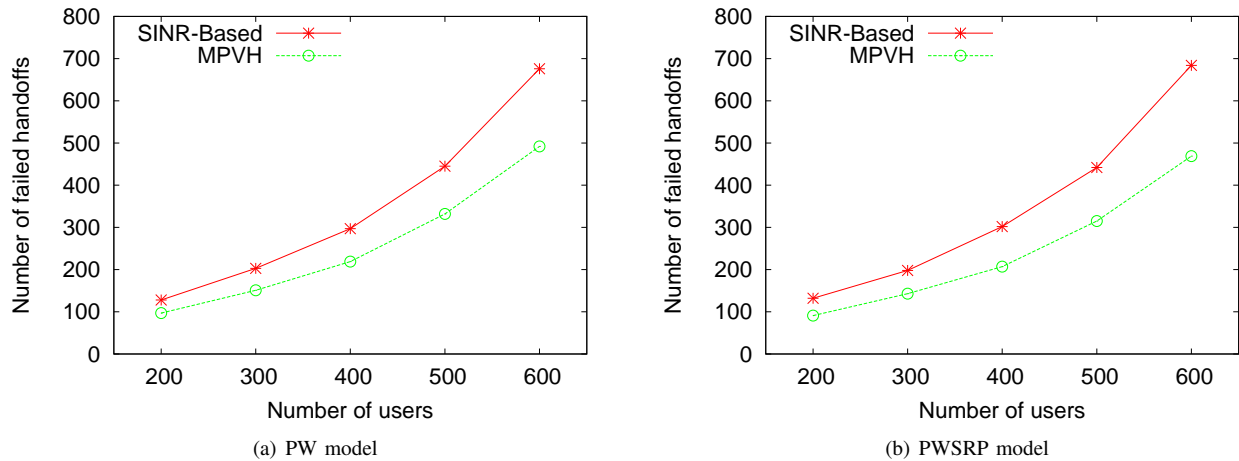


Fig. 4: Comparison of the number of failed handoffs by MPVH and SINR-based schemes under a fixed topology.

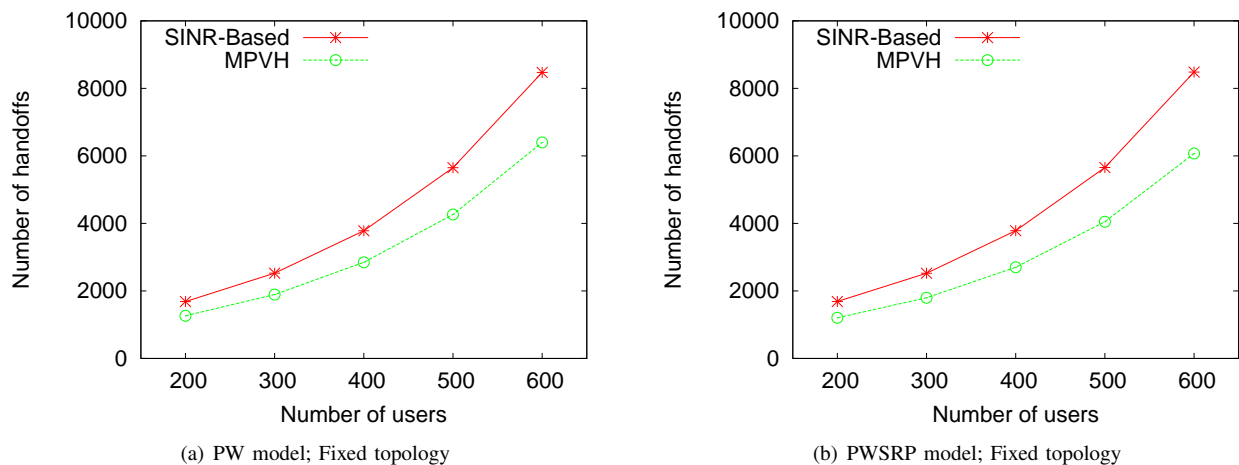


Fig. 5: Comparison of the number of handoffs by MPVH and SINR-based schemes under a fixed topology.

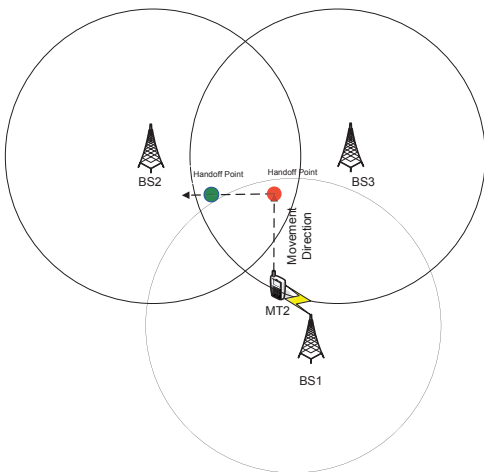


Fig. 6: A case in which MPVH outperforms the SINR-based scheme.

- [8] E. Gustafsson and A. Jonsson. Always best connected. *Wireless Communications, IEEE*, 10(1):49–55, 2003.
- [9] E. Kanan, G. Husari, M. Al-Ayyoub, and Y. Jararweh. Towards improving channel switching in cognitive radio networks. In *Information and Communication Systems (ICICS), 2015 6th International Conference on*, pages 280–285, April 2015.
- [10] M. Lott, M. Siebert, S. Bonjour, D. von Hugo, and M. Weckerle. Interworking of WLAN and 3G systems. *Communications, IEE Proceedings-*, 151(5):507–513, 2004.
- [11] J.G. Markoulidakis, G.L. Lyberopoulos, D.F. Tsirkas, and E.D. Sykas. Mobility modeling in third-generation mobile telecommunications systems. *Personal Communications, IEEE*, 4(4):41–56, 1997.
- [12] S. Mohanty and I.F. Akyildiz. A cross-layer (layer 2+3) handoff management protocol for next-generation wireless systems. *Mobile Computing, IEEE Transactions on*, 5(10):1347–1360, 2006.
- [13] Nouri Omheni, Faouzi Zarai, Mohammad S Obaidat, and Kuei-Fang Hsiao. A novel vertical handoff decision making algorithm across heterogeneous wireless networks. In *Computer, Information and Telecommunication Systems (CITS), 2014 International Conference on*, pages 1–6. IEEE, 2014.
- [14] V Tabataba Vakili and A Aziminejad. A new dynamic channel allocation scheme based on compact pattern concept with pattern restoration. *IRANIAN JOURNAL OF SCIENCE AND TECHNOLOGY*, 28(B3):373–382, 2004.



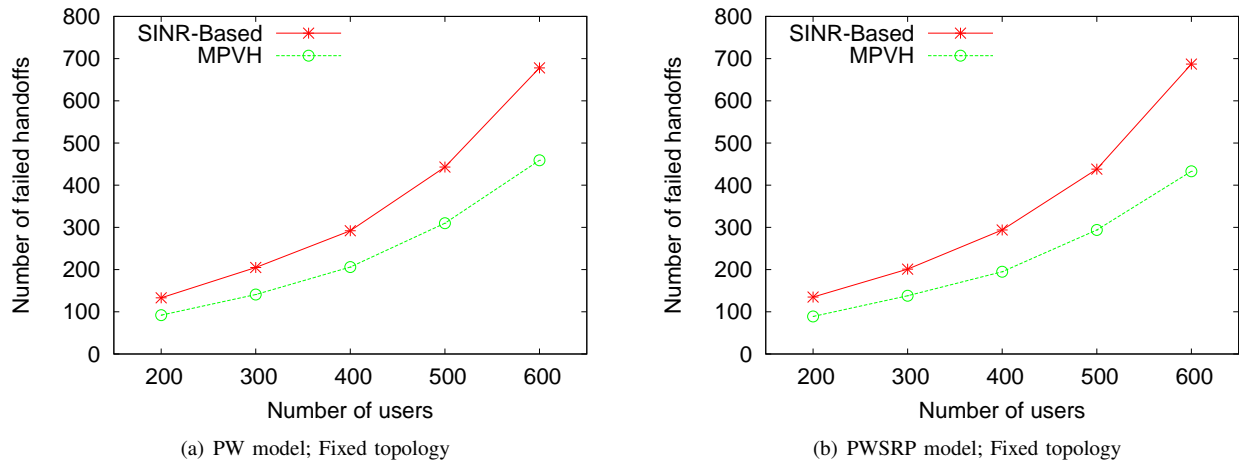


Fig. 7: Comparison of the number of failed handoffs by MPVH and SINR-based schemes under a random topology.

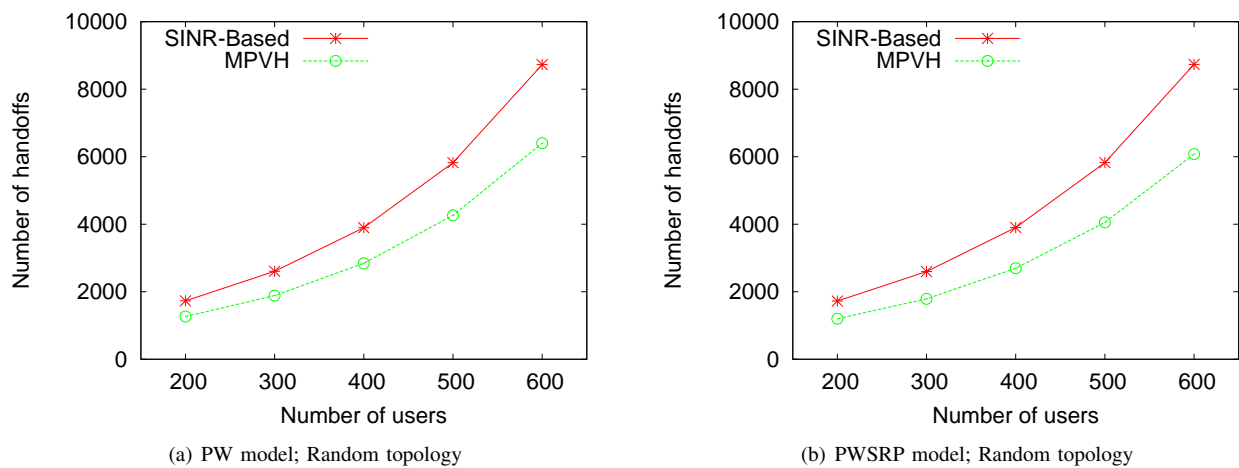


Fig. 8: Comparison of the number of handoffs by MPVH and SINR-based schemes under a random topology.

- [15] Hao Wang, David I Laurenson, and Jane Hillston. A general performance evaluation framework for network selection strategies in 3g-wlan interworking networks. *Mobile Computing, IEEE Transactions on*, 12(5):868–884, 2013.
- [16] T.A. Wysocki, A. Dadej, and B.J. Wysocki. *Advanced wired and wireless networks*, volume 26. Springer, 2004.
- [17] L. Xia, L. Jiang, and C. He. A novel fuzzy logic vertical handoff algorithm with aid of differential prediction and pre-decision method. In *Communications, 2007. ICC'07. IEEE International Conference on*, pages 5665–5670. IEEE, 2007.
- [18] X. Yan, Y. Ahmet Şekercioglu, and S. Narayanan. A survey of vertical handover decision algorithms in fourth generation heterogeneous wireless networks. *Computer Networks*, 54(11):1848–1863, 2010.
- [19] X. Yan, N. Mani, and YA Cekercioglu. A traveling distance prediction based method to minimize unnecessary handovers from cellular networks to WLANs. *Communications Letters, IEEE*, 12(1):14–16, 2008.
- [20] X. Yan, Y.A. Sekercioglu, and N. Mani. A method for minimizing unnecessary handovers in heterogeneous wireless networks. In *World of Wireless, Mobile and Multimedia Networks, 2008. WoWMoM 2008. 2008 International Symposium on a*, pages 1–5. IEEE, 2008.
- [21] Xiaohuan Yan, Y Ahmet Sekercioglu, and Sathya Narayanan. Optimization of vertical handover decision processes for fourth generation heterogeneous wireless networks. In *IWCMC*, pages 1383–1390, 2011.
- [22] K. Yang, I. Gondal, and B. Qiu. multi-dimensional adaptive SINR based vertical handoff for heterogeneous wireless networks. *Communications Letters, IEEE*, 12(6):438–440, 2008.
- [23] K. Yang, I. Gondal, B. Qiu, and L.S. Dooley. Combined SINR based vertical handoff algorithm for next generation heterogeneous wireless networks. In *Global Telecommunications Conference, 2007. GLOBECOM'07. IEEE*, pages 4483–4487. IEEE, 2007.
- [24] A.H. Zahran and B. Liang. Performance evaluation framework for vertical handoff algorithms in heterogeneous networks. In *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, volume 1, pages 173–178. IEEE, 2005.
- [25] A.H. Zahran, B. Liang, and A. Saleh. Signal threshold adaptation for vertical handoff in heterogeneous wireless networks. *Mobile Networks and Applications*, 11(4):625–640, 2006.
- [26] J. Zhang, H.C.B. Chan, and V.C.M. Leung. WLC14-6: A location-based vertical handoff decision algorithm for heterogeneous mobile networks. In *Global Telecommunications Conference, 2006. GLOBECOM'06. IEEE*, pages 1–5. IEEE, 2006.

# Performance Evaluation of Affinity Propagation Approaches on Data Clustering

R. Refianti, A.B. Mutiara, *Member IEEE*, and A.A. Syamsudduha  
Faculty of Computer Science and Information Technology  
Gunadarma University  
Jl. Margonda Raya No.100, Depok 16424, Indonesia

**Abstract**—Classical techniques for clustering, such as k-means clustering, are very sensitive to the initial set of data centers, so it need to be rerun many times in order to obtain an optimal result. A relatively new clustering approach named Affinity Propagation (AP) has been devised to resolve these problems. Although AP seems to be very powerful it still has several issues that need to be improved. In this paper several improvement or development are discussed in , i.e. other four approaches: Adaptive Affinity Propagation, Partition Affinity Propagation, Soft Constraint Affinity propagation, and Fuzzy Statistic Affinity Propagation. and those approaches are be implemented and compared to look for the issues that AP really deal with and need to be improved. According to the testing results, Partition Affinity Propagation is the fastest one among four other approaches. On the other hand Adaptive Affinity Propagation is much more tolerant to errors, it can remove the oscillation when it occurs where the occupance of oscillation will bring the algorithm to fail to converge. Adaptive Affinity propagation is more stable than the other since it can deal with error which the other can not. And Fuzzy Statistic Affinity Propagation can produce smaller number of cluster compared to the other since it produces its own preferences using fuzzy iterative methods.

**Keywords**—Affinity Propagation, Availability, Clustering, Exemplar, Responsibility, Similarity Matrix.

## I. INTRODUCTION

Nowadays, the need of information in many aspects of life is really high. The information need to be delivered fast and accurate. It makes the extraction process of information from data is really crucial. There are many ways in mining an information from data, one of them is clustering. Clustering is commonly used to analyze data which is have very large or even huge data in numbers and the class label on data are unknown. Since assigning class labels to large number of data are very high cost process so then another approach such as clustering is needed to mine useful information from data. The Clustering is focused on finding methods for efficient and effective cluster analysis in large databases [4].

In clustering usually it is need to select some cluster centers to guarantee that the sum of squared errors between each data point and its potential cluster center is small during clustering. Classical techniques for clustering, such as k-means clustering, partition the data into k clusters and are very sensitive to the initial set of data centers, so they often need to be rerun many times in order to obtain a satisfactory result.

A relatively new clustering approach named Affinity Propagation (AP) has been devised to resolve these problems [1],

[2]. The AP clustering method has been shown to be useful for many applications in face images, gene expressions and text summarization. Unlike previous methods, AP simultaneously considers all data points as potential exemplars, and it recursively transmits real-valued messages along edges of the network until a good set of centers is generated. In particular, corresponding clusters gradually emerge in AP.

Since its first appearance it already gain much attention from many researchers. Many research and development related to the affinity propagation had been conducted, such as adaptive affinity propagation [11], relational affinity propagation [9], landmark affinity propagation [10], fuzzy statistic affinity propagation [13] and many more emerged since AP introduced as one of powerful clustering algorithm. Although AP has a better solution in clustering data, but it still has several issues related its performance. So AP is still need and open to great possible improvements.

In order to be able in improving AP performance, a wise step to do it by learning from another AP algorithms that have been proposed. By learning and comparing several algorithm, it is possible to gain more information related to their advantages and disadvantages. Later those advantages can be combined to produce a new and better algorithm while at the same time trying to avoid their disadvantages side.

Problems discussed in this research are restricted as following. This research is only meant to find the cluster centers from data images. The data that being used is 2000 data of facial images. And the approaches of affinity propagation that are implemented are only five approaches, i.e. the affinity propagation itself, adaptive affinity propagation [11], partition affinity propagation [12], soft constraint affinity propagation [8], and Fuzzy statistic affinity propagation [13].

The aims of this research are to look for the advantages and disadvantages of affinity propagation development, in order to be able to make a further research to develop a new algorithm which is better and capable to handle clustering faster and more accurate. The research will greatly help a lot of people who is dealing with a large data to make the information extraction easier and faster. Later it is possible for development a new algorithms that can cluster a large data in faster way.

Our paper is organized as follows: In section II, we presents the relevant theories of clustering and its methods, also affinity propagation and its recent developments. Section III, we explains the analysis of affinity propagation and the

application design, and their experiment and analysis results are then presented and discussed in Section IV.

## II. LITERATURE STUDY

### A. Affinity Propagation

Affinity Propagation is a fast clustering algorithm based on message passing that identifies a set of cluster center that represents the dataset [2]. Since it was proposed and published by Frey and Dueck in a paper Science magazine 2007, it is proven has much better performance and much lower clustering error than existing clustering methods [1]. It can support similarities that are not symmetric or do not satisfy the triangle inequality, and it is deterministic which is its clustering results do not depend on initialization unlike most of clustering methods such as k-means [3]. That means AP has some advantages according to its computational speed, general applicability, and good performance [11].

Affinity Propagation works based on similarities between pairs of data points and simultaneously considers all data points as potential exemplars (cluster centers). AP searches for clusters recursively through an iterative process, real valued messages are transmitted or exchanged between data points until high quality set of exemplars and corresponding cluster emerges. In the iterative process, identified exemplars start from the maximum  $n$  number to fewer exemplars until  $m$  exemplars appear and unchanging for a certain number of iterations (converges). The  $m$  clusters found based on  $m$  exemplars are the clustering solution of AP.

Affinity Propagation takes as input a collection of real valued similarities between data points, where the similarity  $S(i, j)$  indicates how well the data point with index  $j$  is suited to be the exemplar for data point  $i$ . The goal of AP is to minimize the squared error, then each similarity is set to a negative squared error (Euclidean Distance). For data point  $x_i$  and  $x_j$  :

$$S(i, j) = -|x_i - x_j|^2 \quad (1)$$

In Affinity Propagation there are two kinds of message exchanged between data points, they are Responsibility and Availability (see Fig. 2). The responsibility  $R(i, j)$  is sent from data point  $i$  to candidate exemplar point  $j$ . It reflects the accumulated evidence for how well suited point  $j$  is to serve as the exemplar for point  $i$ . The availability  $A(i, j)$  is sent from candidate exemplar point  $j$  to data point  $i$ . It reflect the accumulated evidence for how appropriate it would be for point  $i$  to choose point  $j$  as its exemplar.

Affinity propagation takes as input a real number  $S(k, k)$  for each data point  $k$  so that data points with larger values of  $(S(k, k))$  are more likely to be chosen as exemplars. These values are referred to as Preferences. The number of identified exemplars (number of clusters) is influenced by the values of the input preferences, but also emerges from the message-passing procedure. The preferences should be set to a common value. This value can be varied to produce different numbers of clusters. The shared value could be the median of the input similarities (resulting in a moderate number of clusters) or their minimum (resulting in a small number of clusters).

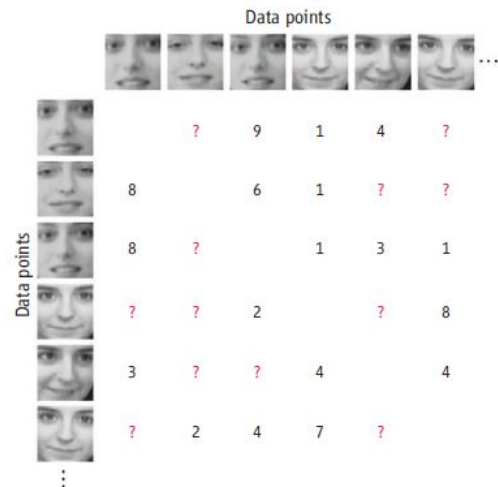


Fig. 1: Similarity Between Data Points [2]

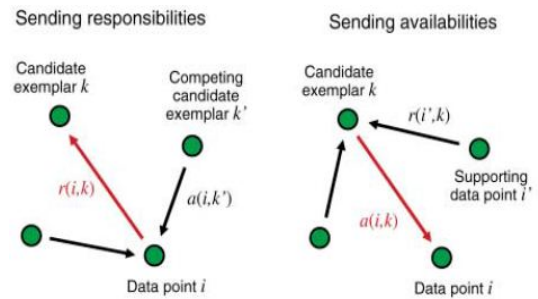


Fig. 2: Responsibility and Availability [2]

When updating the messages, it is important that they be damped to avoid numerical oscillations that arise in some circumstances. Each message is set to  $\lambda$  times its value from the previous iteration plus  $1 - \lambda$  times its prescribed updated value, where the damping factor  $\lambda$  is between 0 and 1.

In each iteration of affinity propagation consisted of updating all responsibilities given the availabilities, updating all availabilities given the responsibilities, and combining availabilities and responsibilities to monitor the exemplar decisions and terminate the algorithm when these decisions did not change for certain number of iterations.

### B. Affinity Propagation Developments

Currently there are many research related to the clustering algorithms of Affinity Propagation. Here will be reviewed several of them. They are Adaptive Affinity Propagation, Partition Affinity Propagation, Soft Constraint Affinity Propagation, Fast Sparse Affinity Propagation [7], Relational Affinity Propagation [9], Fuzzy Statistic Affinity Propagation, Patch Affinity Propagation [15], and Fast Affinity Propagation [3].

1) *Adaptive Affinity Propagation*: Adaptive Affinity Propagation is one of the development of affinity propagation algorithm which is developed based on the fact that the original

affinity propagation is still have several issues. Affinity propagation is known that its clustering result is really influenced by the initial value of the preference parameter (diagonal values of similarity matrix). Minimum value of preference will drive to a small number of cluster center result, while median value will drive to a moderate result. Then how to know what is the best value to be set on the preference to produce the optimal clustering solution. So this algorithm is designed by Kaijun Wang to overcome this issue [11].

Beside the preference issue there is another issue that this algorithm try to deal with. In some cases, the occurrence of oscillations can not be avoided. Oscillation is a condition where the identified exemplar is keep changing in each iteration which is caused the process is fail to reach convergence. And if it made into a graph, the graph will show a periodic graph. The requirement of convergence is that the exemplar is not changing for certain amount of iteration. On original affinity propagation when the oscillation occurs then the clustering process is failing so it is need to be rerun with increasing the damping factor values. But there is a fact that the greater value of damping factor the slower the process will take times.

Adaptive Affinity Propagation divided into three main parts. The first part is called Adaptive Damping which is the process of adjusting the damping factor to eliminate oscillations adaptively when the oscillations occur. The second part is called Adaptive Escape which is the process of escaping the oscillations by decreasing the preferences value when adaptive damping method fails. And the last part is called Adaptive Preference Scanning which is the process of searching the space of preferences to find out the optimal clustering solution to the data set.

2) *Partition Affinity Propagation*: Partition Affinity Propagation is an extension of affinity propagation which can reduce the number of iterations effectively and has the same accuracy as the original affinity Propagation [12]. This extension comes at the cost of making a uniform sampling assumption about the data. Partition affinity propagation passes messages in the subsets of data first and then merges them as the number of initial step of iterations, it can effectively reduce the number of iterations of clustering.

In each iteration of affinity propagation the updating of each responsibility  $R(i, j)$  takes availabilities and similarities between data points into consideration, while the updating of availability  $A(i, j)$  has a relationship with responsibilities between other data points. Therefore, given the similarity collection, the clustering result achieved by AP is the same with the same preference value as input. However, initial availability matrix is also a factor which will lead to different identified exemplar set [14]. To produce different identified exemplar set by changing initial availabilities matrix of AP. This algorithm reduces the time spent by decreasing the number of iterations by decomposing the original similarity matrix into submatrices

3) *Soft Constraint Affinity Propagation*: Soft Constraint Affinity Propagation is one of the development of affinity propagation algorithm which is developed based on the fact that the original affinity propagation is still suffers from a number of drawback. The hard constraint of having exactly one exemplar per cluster restricts AP to classes of regularly shaped clusters, and leads to suboptimal performance [8].

Each point in the cluster on AP refers to exemplar, and each exemplar is required to refer to itself as a self-exemplar. This hard constraint forces clusters to appear as stars of radius one. There is only one central node, and all other nodes are directly connected to it. The hard constraint in AP relies strongly on cluster shape regularity, elongated or irregular multi-dimensional data might have more than one simple cluster center.

This problems may be solved by modifying the original optimization task of AP. This softening of hard constraint is using a finite penalty term for each constraint violation.

4) *Fuzzy Statistic Affinity Propagation*: Fuzzy Statistic Affinity Propagation [13] is one of the development of affinity propagation algorithm which is developed based on fuzzy statistics. Fuzzy statistics is a combination of statistical methods and fuzzy set theory. Fuzzy set theory is the basis in studying membership relationships from the fuzziness of the phenomena. Fuzzy statistics is used to estimate the degree of membership of a pattern to a class according to the use of a membership function. At the Fuzzy Statistic Affinity Propagation compute fuzzy mean deviation and then develop a fuzzy statistical similarity measure (FSS) in evaluating the similarity between two pixel vectors before running the AP iterative process.

### III. RESEARCH METHODOLOGY

#### A. System Analysis

In order to conduct research about clustering by Affinity Propagation, there are several steps that have to be considered and to be done. These steps are data collection, data preprocessing, similarity matrix construction, evidence calculation, and cluster exemplar identification.

1) *Data Collection*: Data collection is the very first step to do in cluster analysis. In this research, data are searched and collected through internet. The data are consists of 250x250 pixels 8500 facial images which is collected from 1500 living people category on Wikipedia [5]. These data has been widely used in several researches such as mining faces biographies [6]. But in this research only 2000 facial images will be used. Here are several samples of the images.



Fig. 3: Sample Data Images [5]

2) *Data Preprocessing*: Before the collected data proceeded to the clustering process, these data are needed to be preprocessed first to eliminate noise and irrelevant information. These steps include image cropping, image resizing, and image gray scaling.

3) *Similarity Matrix Construction*: After the data are completely preprocessed, it is needed to extract the histogram from the images. This histogram will be used and subtracted with another histogram from the other images to measure its similarities. To do that, a distance approach called Negative Euclidean Distance is used to handle it. Below is the formula of negative Euclidean distance.

$$-|x_i - x_j|^2 \quad (2)$$

According to the Euclidean distance the diagonal matrix should be equal to zero, since it is being subtracted from itself. These zero values need to be replaced with other values. These values later will be called as the preference value. This preference value later will influence the clusters output and the number of clusters. It is one of the crucial parameters in affinity propagation. Different values will cause different cluster results. It can be said that the number of identified clusters is increased or decreased by adjusting the preference value correspondingly [11].

There are two options that usually are chosen to set the preference value, by setting it to its median from the similarity matrix  $median(S)$  or the other hand is setting it to its minimum  $min(S)$ . The median value will lead to a moderate number of clusters, while the minimum result in a small number of clusters. Besides these two options, a custom value can also be set if it possibly can give a better cluster solution.

4) *Evidence Calculation*: Evidence calculation is the main part of affinity propagation which is all the computation takes place. It is divided into two main parts. The first one is called Responsibility Update, and the second is Availability Update. The responsibility  $R(i, j)$  reflects the accumulated evidence for how well suited point  $j$  is to serve as the exemplar for point  $i$ . While the availability  $A(i, j)$  reflects the accumulated evidence for how appropriate it would be for point  $i$  to choose point  $j$  as its exemplar.

First will be illustrated how to compute the responsibility update from the similarity matrix. But before that it is needed to be predefined some parameters such as set the maximum iteration number, convergence iteration, and damping factor. For the first iteration make sure that the availability is set to zero  $A(i, j) = 0$ .

The Responsibility Update itself is based on these formulas.

$$R(i, j) = \begin{cases} S(i, j) - \max\{A(i, j) + S(i, j)\} & (i \neq j) \\ S(i, j) - \max\{S(i, j)\} & (i = j) \end{cases} \quad (3)$$

To gain an understanding about how the responsibility update actually works, here is presented the algorithm of Responsibility Update.

To illustrate the above algorithm, the first step is adding the similarity matrix with the initial Availability which is now zero for the first iteration. In the next iteration it will no longer remain zero, some of it will be replaced by negative or positive numbers.

$$AS(i, j) = A(i, j) + S(i, j) \quad (4)$$

---

### Algorithm 1 . Responsibility Update Calculation

---

**Input** : Similarity Matrix  $S(i, j)$ , Availability Matrix  $A(i, j)$ ,  
Maximum Iteration, Convergence Iteration, Damping Factor

**Output** : Responsibility Matrix  $R(i, j)$

- 1: Set Matrix  $R(i, j) = 0$  and  $A(i, j) = 0$  (first iteration only)
  - 2: Compute  $AS(i, j) = A(i, j) + S(i, j)$
  - 3: Compute max value and record the index  $[maxVal(i), Index] = max(AS)$
  - 4: Replace max value with max negative real number/inf  $AS(Index) = -inf$
  - 5: Compute 2nd max value  $[maxVal2(i)] = max(AS)$
  - 6: Make vector max into matrix  $maxAS(i, j) = repmat(maxVal(i))$
  - 7: Replace max Value Index with 2nd max Value  $maxAS(Index) = maxVal2(i)$
  - 8: Subtract S with maxAS  $R(i, j) = S(i, j) - maxAS(i, j)$
  - 9: Compute final  $R(i, j) = (1 - lam) * R(i, j) - lam * R(i, j)$
- 

Then find the maximum values from the matrix  $AS(i, j)$ . And record the index locations. Replace the maximum values with any other value that will always be the greatest value of all values in the matrix. In this case it can be replaced by an infinite number or real maximum number. This is meant to make sure that those index will no longer be chosen as the maximum value in the matrix so it can choose the second maximum values. Since the negative value of infinite number of real maximum number will always become the smallest number. With the replaced maximum values, we find the second maximum values from the matrix.

Make a new matrix with the same size as matrix  $AS(i, j)$  from the first maximum values. Replace some values according to the recorded index from the first maximum values with the second obtained maximum values.

Then subtract the similarity matrix  $S(i, j)$  with the matrix  $maxAS(i, j)$  and then will be obtained a new matrix which is called responsibility matrix  $R(i, j)$ . Below is the subtraction process.

$$R(i, j) = S(i, j) - maxAS(i, j) \quad (5)$$

And the last step is to multiply the obtained matrix with the damping factor. This damping factor will be really useful later to prevent an oscillation which will cause the algorithm to fail to reach a convergence. Convergence itself is the condition where the identified number of clusters is no longer changing for certain number of iterations which is called as convergence iteration. And the formula of damping factor is like follow.

$$R(i, j) = (1 - \lambda) \cdot R(i, j) - \lambda \cdot R(i, j) \quad (6)$$

The damping factor ( $\lambda$ ) value can be set between zero to one. In this illustration the value is set to 0.5. And finally the final responsibility matrix is obtained like below. This matrix will be the input for the availability update.

The second main part of affinity propagation is to calculate the availability matrix which is called Availability Update. The



input of this step is still the same as the previous one plus the new matrix responsibility. Availability update process is based on these formula

$$A(i, j) = \begin{cases} \min(0, (R(j, j) + \sum \max(0, R(i, j)))) & (i \neq j) \\ \sum \max(0, R(i, j)) & (i = j) \end{cases} \quad (7)$$

To gain an understanding about how the availability update actually works, here is presented the algorithm of Availability Update.

---

**Algorithm 2 . Availability Update Calculation**

---

**Input :** Responsibility Matrix  $R(i, j)$

**Output :** Availability Matrix  $A(i, j)$

- 1: Find  $Rp(i, j) = R(i, j) > 0$
  - 2: Put diagonal  $R(i, j)$  to diagonal  $Rp(i, j)$
  - 3: Compute sum of  $Rp(i, j)$
  - 4: Make new matrix from  $sum(Rp(i, j))$
  - 5: Compute matrix  $A(i, j) = sum(Rp(i, j)) - Rp(i, j)$
  - 6: Find  $A(i, j) < 0$
  - 7: Put diagonal  $A(i, j)$  to diagonal  $A(i, j) < 0$
  - 8: Compute final  $A(i, j) = (1 - lam) * A(i, j) - lam * A(i, j)$
- 

To illustrate the algorithm above, here used the previous calculated responsibility matrix as the input. The first step is to find the values within responsibility matrix which is greater than zero. Then put the value of diagonal responsibility matrix to the new matrix. This new matrix then is called as responsibility preference  $Rp(i, j)$  matrix. From this matrix add all values in the same column to produce a vector responsibility.

Make a new matrix from this vector that will be used later to be subtracted by matrix responsibility preference. Subtract this matrix with matrix responsibility preference  $Rp(i, j)$  to produce matrix Availability  $A(i, j)$ . Then from this matrix Availability  $A(i, j)$  look for the values that have smaller values than zero. The values from diagonal of availability matrix is placed to the diagonal of new matrix availability then the final availability matrix will be formed.

And the last is the same like responsibility matrix, multiply it with the damping factor as shown below.

$$A(i, j) = (1 - \lambda) \cdot A(i, j) - \lambda \cdot A(i, j) \quad (8)$$

5) *Cluster Exemplars Identification:* In order to be able to identify an exemplars or cluster centers from the similarity matrix using affinity propagation there are two things that need to be created, they are the responsibility and availability matrix. At the previous explanation, both of responsibility and availability is successfully created. These two matrixes later will determine the clustering result. To identify the cluster centers, it simply just need to add both matrix responsibility and availability like follow.

$$C(i, j) = R(i, j) + A(i, j) \quad (9)$$

From this matrix  $C(i, j)$  the cluster centers easily determined by looking for the value that greater than zero. These whole process from the responsibility update until cluster

identification will be done repeatedly until the number of identified cluster is unchanging for certain amount of times.

### B. Affinity Propagation Analysis

Previously it is already described the whole system of affinity propagation and how it is actually works. Now in this part will be presented some of the developments related to the affinity propagations. They are Adaptive Affinity Propagation, Partition Affinity Propagation, Soft Constraint Affinity Propagation Algorithm, and Fuzzy Statistic Affinity Propagation.

---

**Algorithm 3 . Adaptive Damping and Escape**

---

**Input :** Monitoring Window  $w = 40$ , Damping Factor  $lam = 0.5$ ,

Maximum Iteration  $maxits$ , Preference Step  $P_s$

**Output :** Exemplars of Each Data Points

- 1: **for**  $i = 0$  to  $i = 10$  **do**
  - 2:   Record IdentIFied Exemplars  $Kset(i) = K$
  - 3:   **if**  $i > w/8$  **then**
  - 4:     Compute  $Kmean(i) = mean(Kset(i - (w/8) : i))$
  - 5:     Check FOR Decreasing K
  - 6:     **if**  $Kmean(i) - Kmean(i - 1) < 0$  **then**
  - 7:       K Decreased is true  $Kd = 1$
  - 8:     **end if**
  - 9:     Check FOR Uchanging K
  - 10:    **if**  $Kset(i) - Kset(i - 1) == 0$  **then**
  - 11:      K changed is True  $Kc = 1$
  - 12:    **end if**
  - 13:    Check FOR non-oscillation
  - 14:    **if**  $Kd == 1$  or  $Kc == 0$  **then**
  - 15:       $Kb(i) = 1$
  - 16:    **end if**
  - 17:    Check FOR Oscillation Occurance
  - 18:    **if**  $sum(Kb) < 2/3w$  **then**
  - 19:      Activate Adaptive Damping
  - 20:      Increase damping factor  $lam + 0.05$
  - 21:      **if**  $lam > 0.85$  **then**
  - 22:        Activate Adaptive Escape
  - 23:        Icrease Preference  $P + P_s$
  - 24:      **end if**
  - 25:    **end if**
  - 26:    **end if**
  - 27: **end for**
- 

1) *Adaptive Affinity Propagation Algorithm:* In adaptive affinity propagation proposed a method called Adaptive Damping which is the process of adjusting the damping factor to eliminate oscillations adaptively when the oscillations occur. The second part is called Adaptive Escape which is the process of escaping the oscillations by decreasing the preferences value when adaptive damping method fails.

In adaptive damping method when the oscillations occur, the damping factor will be increased automatically or adaptively once by a step such as 0.05. But after increasing the damping factor until certain number such as 0.85 and the oscillations is still occur then it is used adaptive escape method to deal with it. The adaptive escape is activated if the adaptive damping fail to handle the oscillation by increasing the preference values of similarity matrix until the process reach convergence.

---

**Algorithm 4 .** Partition Affinity Propagation

---

**Input :** Number of Data Point  $N$ , Similarity Matrix  $S(i, j)$   
**Output :** Exemplars of Each Data Points

- 1: Calculate Number of Partition  $N_{part} = \text{sqrt}(N)/2$
- 2: Calculate Matrix Partition Size  $PartSize = N/N_{part}$
- 3: Set Partition End point to zero  $EndP = 0$
- 4: **for**  $i = 1$  to  $N_{part}$  **do**
- 5:   Calculate Partition Start Point  $StartP = 1 + PartSize * (i - 1)$
- 6:   **if**  $i == N_{part}$  **then**
- 7:     Calculate Matrix Size  $PartSize = N - PartSize * (N_{part} - 1)$
- 8:     Calculate Partition End Point  $EndP = EndP + PartSize$
- 9:   **else**
- 10:    Calculate Partition End Point  $EndP = EndP + PartSize$
- 11:   **end if**
- 12:   Construct Sub Matrix  $SM(i, j) = S(StartP : EndP, StartP : EndP)$
- 13:   Run Responsibility Update Once with matrix  $SM(i, j)$
- 14:   Run Availability Update Once and save in  $SubA(i)$
- 15: **end for**
- 16: Combine All obtained  $A(i, j) = SubA(*)$
- 17: Run Affinity Propagation with New  $A(i, j)$

---



---

**Algorithm 5 .** Soft Constraint Affinity Propagation

---

**Input :** Similarity Matrix  $S(i, j)$ , Soft Constraint Penalty  $P$   
**Output :** Exemplars of Each Data Points

- 1: Compute Responsibility Update
- 2: Compare diagonal value of matrix  $R(i, i) = \max(-P, R(i, i))$
- 3: Compute Availability Update
- 4: Compare diagonal value of matrix  $A(i, i) = \min(P, A(i, i))$

---

2) *Partition Affinity Propagation Algorithm:* Partition affinity propagation is an extension of affinity propagation which can reduce the time spent to find the exemplar by decreasing the number of iterations. It can be achieved by decomposing the original similarity matrix into sub-matrices. This algorithm passes messages in the subsets of data first and then merges them as the number of initial step of iterations, it can effectively reduce the number of iterations of clustering. The Algorithm is stated as follow.

3) *Soft Constraint Affinity Propagation Algorithm:* Soft Constraint Affinity Propagation is one of the development of affinity propagation algorithm which is developed based on the fact that the original affinity propagation is still suffers from a number of drawback. Each point in the cluster on AP refers to exemplar, and each exemplar is required to refer to itself as a self-exemplar. This hard constraint forces clusters to appear as stars of radius one. This problems may be solved by modifying the original optimization task of AP. This softening of hard constraint is using a finite penalty term for each constraint violation. The Algorithm is stated as follow.

4) *Fuzzy Statistic Affinity Propagation Algorithm:* Fuzzy Statistic affinity propagation is an extension of affinity prop-

---

**Algorithm 6 .** Fuzzy Statistic Affinity Propagation

---

**Input :** Number of Data  $N$ , Data Points  $X$ , exemplar  $Z$ , membership degree  $U$ , degree of fuzziness  $B$ , error rate  $E$ , Cluster Threshold Scalar  $CTS$   
**Output :** Similarity Matrix  $S(i, j)$ , and Preference  $P$

- 1: Set values of  $B$ ,  $E$ , and  $CTS$
- 2: Set initial  $U = 1$ , fuzzy mean distance  $fmdist = 0$ , and stopping condition  $stop = 0$
- 3: Calculate fuzzy distance  $fdist = |X - Z|$
- 4: **while**  $stop == 0$  **do**
- 5:   Record old fuzzy mean distance  $fmdist1 = fmdist$
- 6:   Calculate fuzzy mean distance  $fmdist = (fdist * U) / U$
- 7:   Calculate fuzzy distance deviation  $fdev = fdist - fmdist$
- 8:   Calculate fuzzy mean distance deviation  $fmdev = (fdist - fmdist) / N$
- 9:   Calculate membership degree  $U = \exp((-fdev^B) / (fmdev^B))$
- 10:   Calculate stopping condition  $sc = fmdist1 - fmdist$
- 11:   **if**  $sc < E$  **then**
- 12:      $stop = 1$
- 13:   **end if**
- 14: **end while**
- 15: Calculate fuzzy similarity matrix  $S(i, j) = -fmdist$
- 16: Calculate preference  $P = \min(S(i, j)) - CTS(\max(S(i, j)) - \min(S(i, j)))$

---

agation based on fuzzy statistic and AP. It simultaneously considers all data points in the feature space to be initial clustering exemplars and iteratively refines with the mean distance deviation until getting the optimal fuzzy statistical similarity matrix. The Algorithm is stated as follow.

#### IV. RESULTS AND DISCUSSIONS

Affinity Propagation algorithms is implemented in MATLAB programming environment. Data Preprocessing is a process to eliminate noise and irrelevant information from data. The steps include image cropping, image resizing, and image gray scaling.

The first test result of affinity propagation which is the original algorithm of affinity propagation. It shows that the number of identified cluster centers or exemplars are 6 exemplars from 2000 images with 134 iteration in 23.9285 seconds. From the adaptive affinity propagation algorithm is identified the number of cluster centers or exemplars from 2000 images are 7 exemplars with 247 iteration in 43.3825 seconds. Below is the detailed information including graph and log application.

From the partition affinity propagation algorithm is identified the number of cluster centers or exemplars from 2000 images are 7 exemplars with 119 iteration in 21.0428 seconds. The soft constraint affinity propagation seems like not giving any result since it is fail to meet a convergence. Below is the detailed information including graph and log application.

From the fuzzy statistic affinity propagation algorithm is identified the number of cluster centers or exemplars from 2000 images are 3 exemplars with 64 iteration in 13.3443



TABLE I: Log Minimum Result of Original and Adaptive AP

	Original AP	Adaptive AP
Number of Data	2000	200
Total Iteration	134	249
Total Identifier. Cluster	6	7
Cluster Exemplar	559, 900, 1069, 1074, 1304, 1555	173, 213, 606,769, 1369, 1633, 1859
Cluster Exemplar Index	900, 1304, 900, 1304, 559, 559	1369, 173, 1369, 213, 608, 1633, 769
Execution Time	23.9265 sec.	43.3825 sec.

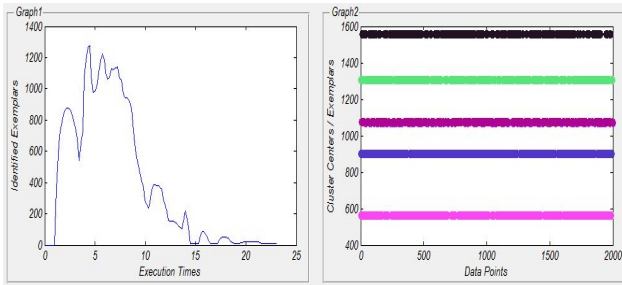


Fig. 4: Graph Minimum Original Affinity Propagation Result

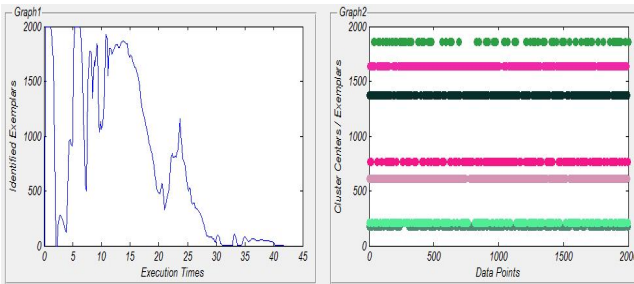


Fig. 5: Graph Minimum Adaptive Affinity Propagation Result

TABLE II: Log Minimum Result of Partition and Soft-Constraint AP

	Partition AP	SC AP
Number of Data	2000	2000
Total Iteration	119	179
Total Identifier. Cluster	7	-
Cluster Exemplar	380, 559, 1069, 1074, 1095, 1497 1555	—
Cluster Exemplar Index	1095, 380, 1095, 380, 559, 559	-
Execution Time	23.9265 sec.	—

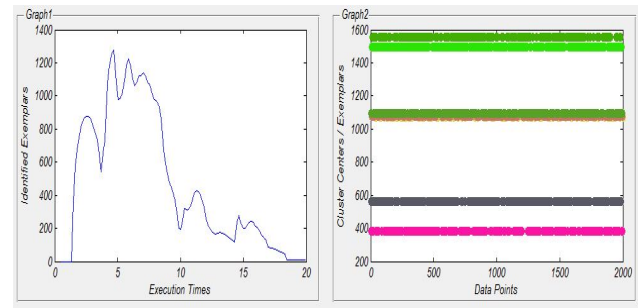


Fig. 6: Graph Minimum Partition Affinity Propagation Result

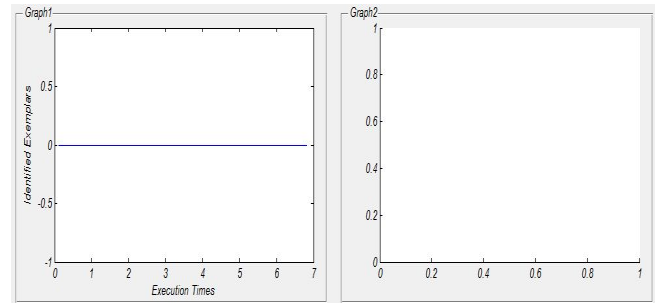


Fig. 7: Graph Minimum Soft Constraint Affinity Propagation Result

seconds. Below is the detailed information including graph and log application.

TABLE III: Log Minimum Result of Fuzzy-Statistics AP

	Fuzzy-Statistic AP
Number of Data	2000
Total Iteration	64
Total Identifier. Cluster	3
Cluster Exemplar	500, 1259, 1770,
Cluster Exemplar Index	1297, 1297, 1297, 500, 1770
Execution Time	13.3443 sec.

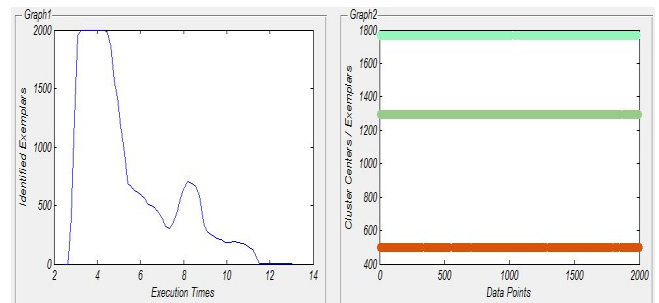


Fig. 8: Graph Minimum Fuzzy Statistic Affinity Propagation Result

Those are the result from the first to which is trying to look

for exemplars from minimum values of preference. The second test result is about to show the identified exemplars from the median values of preferences. the original algorithm of affinity propagation. It shows that the number of identified cluster centers or exemplars are 21 exemplars from 2000 images with 54 iteration in 9.6316 seconds. From the adaptive affinity propagation algorithm is identified the number of cluster centers or exemplars from 2000 images are 22 exemplars with 134 iteration in 23.0912 seconds. Below is the detailed information including graph and log application.

TABLE IV: Log Median Result of Original and Adaptive AP

	Original AP	Adaptive AP
Number of Data	2000	200
Total Iteration	54	134
Total Identifier. Cluster	21	22
Cluster Exemplar	22, 88, 267, 443, 501, 66	144, 460, 537, 586, 634, 672, 685, 725
Cluster Exemplar Index	443, 1333, 1019, 88,	685, 634, 799, 537, 1591, 672, 1729
Execution Time	9.6316 sec.	23.0912 sec.

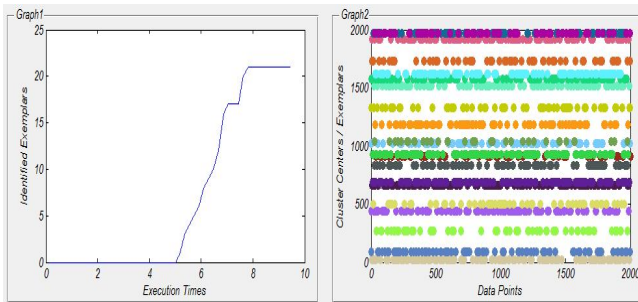


Fig. 9: Graph Median Original Affinity Propagation Result

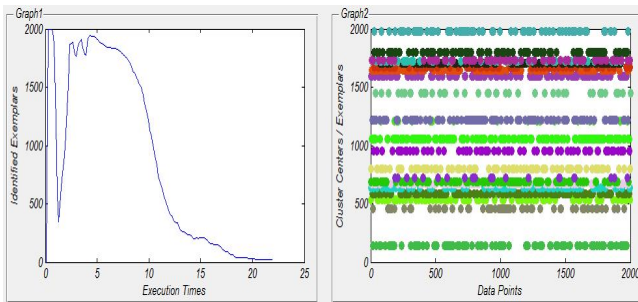


Fig. 10: Graph Median Adaptive Affinity Propagation Result

From the partition affinity propagation algorithm is identified the number of cluster centers or exemplars from 2000 images are 21 exemplars with 57 iteration in 9.5211 seconds.

From the soft constraint affinity propagation algorithm is identified the number of cluster centers or exemplars from 2000 images are 21 exemplars with 54 iteration in 9.9587 seconds. Below is the detailed information including graph and log application.

TABLE V: Log Median Result of Partition and Soft-Constraint AP

	Partition AP	SC AP
Number of Data	2000	2000
Total Iteration	57	54
Total Identifier. Cluster	21	21
Cluster Exemplar	5, 19, 88, 267, 443, 549 690, 834	22, 88, 267 267, 443 501
Cluster Exemplar Index	443, 1333 1019, 8, 5, 1919	443, 1333, 1019, 88
Execution Time	9.5211 sec.	9.9587 sec.

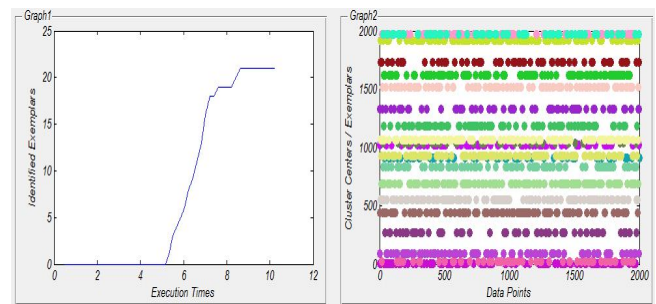


Fig. 11: Graph Median Partition Affinity Propagation Result

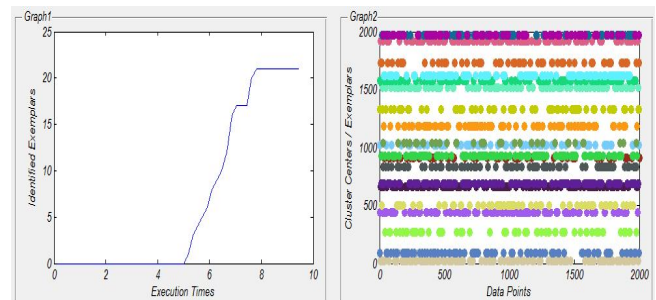


Fig. 12: Graph Median Soft Constraint Affinity Propagation Result

From the fuzzy statistic affinity propagation algorithm is identified the number of cluster centers or exemplars from 2000 images are 3 exemplars with 64 iteration in 13.3443 seconds. Below is the detailed information including graph and log application.

To make it easier to compare the clustering result using affinity propagation approaches. We present Table VII and Table VIII that consist of all the testing results.

TABLE VI: Log Median Result of Fuzzy-Statistics AP

	Fuzzy-Statistic AP
Number of Data	2000
Total Iteration	84
Total Identifier. Cluster	3
Cluster Exemplar	1159, 1671, 1734,
Cluster Exemplar Index	1734, 1734, 1734, 1159, 1671
Execution Time	16.868 sec.

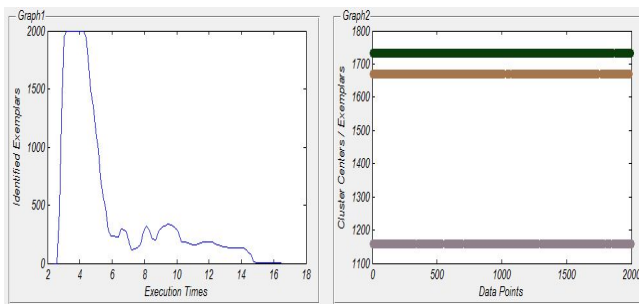


Fig. 13: Graph Median Fuzzy Statistic Affinity Propagation Result

TABLE VII: Testing Result with Minimum Preferences

Algorithms	Minimum Preferences		
	Exemplar	Iteration	Time(s)
Original	6	134	23.9285
Adaptive	7	249	43.3825
Partition	7	119	21.0428
Soft Constraint	-	-	-
Fuzzy Statistic	3	64	13.3443

TABLE VIII: Testing Result with Median Preferences

Algorithms	Median Preferences		
	Exemplar	Iteration	Time(s)
Original	21	54	9.6316
Adaptive	22	134	23.0912
Partition	21	57	9.5211
Soft Constraint	21	54	9.9587
Fuzzy Statistic	3	84	16.868

From the above table original affinity propagation with minimum preferences value generates 6 cluster centers within 23.93 second after 134 iterations, and 21 cluster centers in 9.63 second with 54 iteration for the median preferences. The second algorithm, Adaptive affinity propagation takes much longer time to find the exemplars even compared to the other four approaches it is the slowest algorithm. Slow execution time is the cost of adaptive affinity propagation that has to take since its main goal is to eliminate the possible error occurrence. In some cases in the middle of iteration oscillation will occur and causing the algorithm to fail to converge and need to rerun with different parameters value. Adaptive affinity propagation tries to eliminate this error with sacrificing its speed but more

certain in finding cluster centers. From the table the fastest algorithm is the partition affinity propagation that has the fastest execution time 21.04 second for minimum preference and 9.52 second for median preference compared to the other approaches. Partition affinity propagation has a goal to reduce the iteration numbers so the execution time will be faster. Compared to the original affinity propagation which takes 134 iterations to find 6 exemplars, partition affinity propagation only takes 119 iterations to find 7 exemplars in minimum preferences. and partition affinity propagation takes 57 iterations to find 21 exemplars. Soft constraint affinity propagation with the median preference finds 21 exemplars with 54 iterations, the same as the original affinity propagation but slightly slower execution time since in soft constraint affinity propagation in each iteration it is comparing the diagonal matrix with its soft constraint preference (scp). But in the minimum preference the algorithm fails to converge, it can be caused by the occurrence of oscillations or the value of scp is not well set. The last algorithm, fuzzy statistic affinity propagation resulting in the smallest number of cluster centers or exemplars. It finds 3 exemplars in both minimum and median preference since it generates its own preference using the fuzzy iterative methods. and in the median preference original affinity propagation takes 54 iterations

## V. CONCLUDING REMARKS

According to the testing results, it can be inferred that among several affinity propagation development which is explained in this research Partition Affinity Propagation is the fastest one among four other approaches. Partition Affinity Propagation is slightly faster than the other since it uses an approach to divide matrix into several parts to reduce the iteration numbers. On the other hand Adaptive Affinity Propagation is much more tolerant to errors, it can remove the oscillation when it occurs where the occurrence of oscillation will bring the algorithm to fail to converge. Adaptive Affinity propagation is more stable than the other since it can deal with error which the other cannot. And Fuzzy Statistic Affinity Propagation can produce a smaller number of clusters compared to the other since it produces its own preferences using fuzzy iterative methods.

For future work, in order to produce a better algorithm of affinity propagation consider to try combining two or three algorithms compared in this research such as adaptive and partition affinity propagation or any other affinity propagation. So it can run faster and stabler in generating cluster center from massive data.

## ACKNOWLEDGMENT

The authors would like to thank the Gunadarma Foundations for financial supports.

## REFERENCES

- [1] D. Dueck, "Affinity propagation: clustering data by passing messages," Ph.D. dissertation, Citeseer, 2009.
- [2] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.
- [3] Y. Fujiwara, G. Irie, and T. Kitahara, "Fast algorithm for affinity propagation," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 3. Citeseer, 2011, p. 2238.

- [4] J. Han, M. Kamber, and J. Pei, *Data mining, southeast asia edition: Concepts and techniques*. Morgan kaufmann, 2006.
- [5] M. K. Hasan and C. Pal, "Experiments on visual information extraction with the faces of wikipedia," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
- [6] M. K. Hasan and C. J. Pal, "Improving alignment of faces for recognition," in *Robotic and Sensors Environments (ROSE), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 249–254.
- [7] Y. Jia, J. Wang, C. Zhang, and X.-S. Hua, "Finding image exemplars using fast sparse affinity propagation," in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 639–642.
- [8] M. Leone, M. Weigt *et al.*, "Clustering by soft-constraint affinity propagation: applications to gene-expression data," *Bioinformatics*, vol. 23, no. 20, pp. 2708–2715, 2007.
- [9] A. Plangprasopchok, K. Lerman, and L. Getoor, "Integrating structured metadata with relational affinity propagation?" in *Statistical Relational Artificial Intelligence*, 2010.
- [10] P. Thavikulwat, "Affinity propagation: A clustering algorithm for computer-assisted business simulations and experiential exercises," *Developments in Business Simulation and Experiential Learning*, vol. 35, 2014.
- [11] K. Wang, J. Zhang, D. Li, X. Zhang, and T. Guo, "Adaptive affinity propagation clustering," *arXiv preprint arXiv:0805.1096*, 2008.
- [12] D.-y. Xia, F. Wu, X.-q. Zhang, and Y.-t. Zhuang, "Local and global approaches of affinity propagation clustering for large scale data," *Journal of Zhejiang University SCIENCE A*, vol. 9, no. 10, pp. 1373–1381, 2008.
- [13] C. Yang, L. Bruzzone, F. Sun, L. Lu, R. Guan, and Y. Liang, "A fuzzy-statistics-based affinity propagation technique for clustering in multi-spectral images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 6, pp. 2647–2659, 2010.
- [14] X. Zhang, F. Wu, and Y. Zhuang, "Clustering by evidence accumulation on affinity propagation," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.
- [15] X. Zhu and B. Hammer, "Patch affinity propagation," *ICOLE 2010*, p. 63, 2011.

# Modified Grapheme Encoding and Phonemic Rule to Improve PNNR-Based Indonesian G2P

Suyanto

Department of Computer  
Science and Electronics, FMIPA  
Gadjah Mada University  
Yogyakarta, Indonesia 55281

Sri Hartati

Department of Computer  
Science and Electronics, FMIPA  
Gadjah Mada University  
Yogyakarta, Indonesia 55281

Agus Harjoko

Department of Computer  
Science and Electronics, FMIPA  
Gadjah Mada University  
Yogyakarta, Indonesia 55281

**Abstract**—A grapheme-to-phoneme conversion (G2P) is very important in both speech recognition and synthesis. The existing Indonesian G2P based on pseudo nearest neighbour rule (PNNR) has two drawbacks: the grapheme encoding does not adapt all Indonesian phonemic rules and the PNNR should select a best phoneme from all possible conversions even though they can be filtered by some phonemic rules. In this paper, a modified partial orthogonal binary grapheme encoding and a phonemic-based rule are proposed to improve the performance of PNNR-based Indonesian G2P. Evaluating on 5-fold cross-validation, contain 40K words to develop the model and 10K words to evaluation each, shows that both proposed concepts reduce the relative phoneme error rate (PER) by 13.07%. A more detail analysis shows the most errors are from grapheme ⟨e⟩ that can be dynamically converted into either /ɛ/ or /ə/ since four prefixes, 'ber', 'me', 'per', and 'ter', produce many ambiguous conversions with basic words and also from some similar compound words with both different pronunciations for the grapheme ⟨e⟩. A stemming procedure can be applied to reduce those errors.

**Keywords**—Modified grapheme encoding; phonemic rule; Indonesian grapheme-to-phoneme conversion; pseudo nearest neighbour rule

## I. INTRODUCTION

A phonemization or letter-to-sound conversion, more commonly known as grapheme-to-phoneme conversion (G2P), is an important module in both speech recognition and speech synthesis. In general, a G2P is developed using machine learning-based methods, such as instance-based learning [1], table lookup with defaults [1], self-learning techniques [2], hidden Markov model [3], morphology and phoneme history [4], joint multigram models [5], conditional random fields [6], Kullback-Leibler divergence-based hidden Markov model [7]. These methods are commonly very complex and designed to be language independent, but they give varying performances for some phonemically complex languages, such as English, Dutch, French, and Germany.

In [8], an information gain tree (IG-tree) with best guest strategy is proposed to develop a specific G2P for Indonesian language. It gives PER of 0.99% for training set of 9K words and testing set of 1K unseen words. Another Indonesian G2P developed using PNNR [9] produces slightly higher PER than the IG-tree, around 1.07, but it is capable of disambiguating homograph words. The PNNR-based G2P has two drawbacks. Firstly, the designed grapheme encoding does not adapt all

Indonesian phonemic rules. Secondly, the PNNR should select a best phoneme from all possible conversions even though they can be filtered using some phonemic rules.

When compare to English, the Indonesian language has a much simpler phonemization rule. An initial study on 50K words, collected from the great dictionary of the Indonesian language (*Kamus Besar Bahasa Indonesia Pusat Bahasa* KBBI) third edition, released in 2008, developed by *Pusat Bahasa*, shows that it has 26 graphemes, where 15 graphemes (57%) are pronounced as certain phonemes without any exception and the rest 11 graphemes (43%) are pronounced as some possible phonemes with or without a quite common phonemic rule, as listed in table I. In this research, all graphemes are converted into single-phonemic symbol (SPS) to simplify the alignment process. The SPS are deliberately developed using characters available in any computer keyboard to simplify the implementation. But, in this paper any phoneme is written using the IPA symbol.

In table I, the grapheme ⟨a⟩ can be pronounced as four possible phonemes: 1) /a/, generally if it is followed by graphemes those commonly pronounced as consonants, such as a word 'abad' (century) that is pronounced as /abad/; 2) /aɪ/, usually if it is followed by ⟨i⟩ or ⟨y⟩, such as 'abai' (do not care) that is pronounced as /abai/; 3) a diphthong /aʊ/, commonly if it is followed by ⟨u⟩ or ⟨w⟩, such as 'harimau' (tiger) that is pronounced as /harimau/; and 4) /a+ʔ/, commonly if it is followed by ⟨a⟩, ⟨e⟩, ⟨i⟩, ⟨o⟩, or ⟨u⟩, such as in 'saat' (sometime) that is pronounced as /saʔat/, and 'bait' (couplet) that is pronounced as /baʔit/. The grapheme ⟨a⟩ is not possible to be pronounced as /aɪ/ if it is not followed by ⟨i⟩ nor ⟨y⟩. It also not possible to be pronounced as /aʊ/ if it is not followed by ⟨u⟩ nor ⟨w⟩. Such phonemic rules can be actually used to filter possible conversions so that PNNR can convert a grapheme into a correct phoneme more accurately and faster.

The grapheme ⟨a⟩ is much more frequently pronounced as /a/, up to 54K, among the three other phonemes those are less than 1K. This fact makes ⟨a⟩ can be converted into a correct phoneme easily. But, grapheme ⟨e⟩ can be pronounced as five possible phonemes: /ɛ/, /ə/, /eɪ/, /ɛ+ʔ/, and /ə+ʔ/. The biggest challenge is converting this grapheme into /ɛ/ and /ə/ since it changes so dynamic without certain phonemic rule and they have high frequencies and many ambiguities. For

TABLE I: Twenty Six Indonesian Graphemes, Their Possible Pronunciations in the Single Phonemic Symbol (SPS) and the International Phonemic Alphabet (IPA) as well as Frequencies and Percentages in 50K words containing 385K Graphemes

Number	Grapheme	SPS	IPA	Frequency	Percentage
1	a	a	ɑ	54,859	14.23%
1	a	A	aɪ	979	0.25%
1	a	U	au	624	0.16%
1	a	l	ɑ+ʔ	669	0.17%
2	b	b	b	11,236	2.91%
3	c	c	tʃ	3,667	0.95%
4	d	d	d	8,077	2.10%
5	e	e	ɛ	9,851	2.56%
5	e	E	ə	30,554	7.93%
5	e	Y	eɪ	29	0.01%
5	e	2	ɛ+ʔ	36	0.01%
5	e	3	ə+ʔ	193	0.05%
6	f	f	f	2,114	0.55%
7	g	g	g	6,492	1.68%
7	g	*	*	11,513	2.99%
8	h	h	h	5,769	1.50%
8	h	*	*	245	0.06%
9	i	i	i	26,685	6.92%
9	i	*	*	1,047	0.27%
9	i	4	i+ʔ	30	0.01%
10	j	j	tʃ	3,381	0.88%
11	k	k	k	21,784	5.65%
11	k	K	x	217	0.06%
11	k	*	*	19	0.00%
12	l	l	l	15,465	4.01%
13	m	m	m	19,237	4.99%
14	n	n	n	22,143	5.74%
14	n	G	ŋ	11,779	3.06%
14	n	N	ɲ	3,741	0.97%
15	o	o	ɔ	13,763	3.57%
15	o	O	əɪ	56	0.01%
15	o	5	ɔ+ʔ	60	0.02%
16	p	p	p	12,919	3.35%
17	q	k	k	23	0.01%
18	r	r	r	24,709	6.41%
19	s	s	s	17,602	4.57%
19	s	S	ʃ	206	0.05%
20	t	t	t	18,981	4.92%
21	u	u	u	17,926	4.65%
21	u	*	*	623	0.16%
21	u	6	u+ʔ	19	0.00%
22	v	f	f	745	0.19%
23	w	w	w	1,784	0.46%
24	x	s	s	29	0.01%
25	y	y	j	1,124	0.29%
25	y	*	*	2,085	0.54%
26	z	z	z	397	0.10%

example, a word 'reses' (recess) is pronounced as /rɛsɛs/, but 'resesi' (recession) is pronounced as /rɛsɛsi/. A grapheme ⟨e⟩ in 'berang' (irascible) is converted into /ɛ/, but ⟨e⟩ in 'berangin' (windy) should be pronounced as /ə/. Therefore, the grapheme ⟨e⟩ is predicted to produce so many errors.

Those facts motivate this research to focus on modifying the partial orthogonal binary grapheme encoding used in [9] and incorporating some phonemic rules to reduce the PER of the PNNR-based Indonesian G2P as well as do a more detail evaluation to see its drawbacks.

In the following sections, this paper will discuss how to use PNNR to develop the Indonesian G2P, the proposed modified partial orthogonal binary grapheme encoding and the phonemic rule-based phoneme filtering, the experimental results showing the performance of both proposed concepts, and the conclusion.

Grapheme sequence :  
<abai>

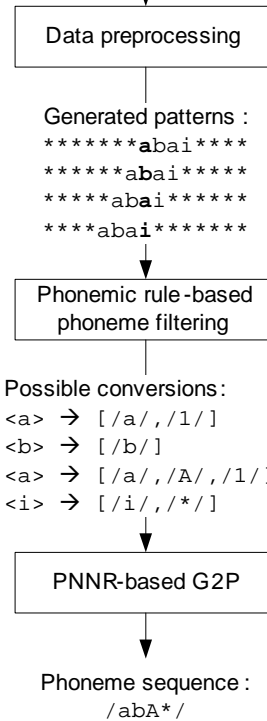


Fig. 1: Block Diagram of Indonesian G2P Using PNNR and Phonemic Rule

## II. PNNR-BASED G2P

The block diagram of PNNR-Based G2P is illustrated by figure 1. Data preprocessing converts a word (grapheme sequence) into some patterns, where '\*' in generated patterns is a symbol for no phoneme and /\*/ in possible conversions is a symbol for no phoneme, while /a/, /1/, and /A/ are single phonemic symbols for /a/, /a+ʔ/, and /ai/ respectively. The phonemic rule filters some potential conversions to be selected by PNNR, for instance the first grapheme ⟨a⟩ followed by ⟨b⟩ in the given grapheme sequence ⟨abai⟩ is possible to be converted into either /a/ or /a+ʔ/. Finally, the PNNR decides the best conversion of each given grapheme into the possible phonemes.

### A. Data Preprocessing

The data used to develop a PNNR-based G2P is a pair of word (sequence of graphemic symbols) and the corresponding pronunciation (sequence of phonemic symbols). The data preprocessing is the sama as described in [9]. First, each grapheme should be mapped into a single phonemic symbol (SPS), as listed in table I. Next, each word should be aligned to the corresponding phonemic symbols. Then, each grapheme occurring in the word is consecutively located as the focus grapheme and the rests on their appropriate contextual positions using contextual length  $L = 14$  (7 graphemes on the left and 7 on the right of the focus, as suggested in [9]) as illustrated by figure 1.



### B. Modified Grapheme Encoding

The partial orthogonal binary grapheme encoding described in [9] has two problems. Firstly, the distance between a vowel-oriented grapheme and a consonant-oriented grapheme is exactly same as the distance between two consonant-oriented graphemes come from different groups or between a consonant-oriented grapheme and a non-grapheme. This encoding, in some cases, makes the PNNR do a wrong conversion. Secondly, three graphemes ⟨q⟩, ⟨v⟩, and ⟨x⟩ are encoded into three unique group although, in fact, those graphemes are always converted into /k/, /f/, and /s/ respectively.

Based on those facts, the partial orthogonal binary grapheme encoding in [9] is modified as follow:

- 1) All graphemes and non-graphemes are divided into three main categories: vowel-oriented graphemes, consonant-oriented graphemes, and non-graphemes. Vowels and consonants are designed to have 6 different bits since they are contextually so different in a word (sequence of grapheme). For example, a grapheme ⟨a⟩ followed by a vowel ⟨i⟩ in 'pantai' (beach) should be pronounced as /ai/, but it should be converted into /a/ when it is followed by a consonant ⟨s⟩ 'pantas' (feasible). But, vowels and non-graphemes as well as consonants and non-graphemes have 5 different bits because they have relatively lower differences. For example, a non-graphemic symbol ⟨-⟩ in 'berang-berang' (beaver) make grapheme ⟨e⟩ pronounced as /ə/, but grapheme ⟨e⟩ in 'berang' (irascible) and 'memberang' (become angry) should be pronounced as /ɛ/.
- 2) In each main category, some small groups developed by considering the pronunciation similarity of those graphemes in manner and place of articulation as described in [10], [11]. Graphemes in a same group are designed to have 2 different bits since they are contextually so similar in a word, for instance a grapheme ⟨n⟩ followed by either ⟨g⟩ or ⟨k⟩ should be converted into /ŋ/ such as 'bang' (brother) and 'bank' (bank) those pronounced as /baŋ/. Graphemes in different groups are designed to have 4 different bits as they are contextually quite different, for instance a grapheme ⟨e⟩ preceded by ⟨p⟩ should be converted into /ə/ such as in word 'peran' (role) that pronounced as /pəran/, but it should be converted into /ɛ/ when it is preceded by ⟨s⟩ such as in 'seran' (strip) that pronounced as /seran/. Therefore, graphemes ⟨k⟩ and ⟨g⟩ are grouped into a group, but ⟨p⟩ and ⟨s⟩ are in a different group.
- 3) Three graphemes ⟨q⟩, ⟨v⟩, and ⟨x⟩ are encoded into the same binary code as graphemes ⟨k⟩, ⟨f⟩, and ⟨s⟩ since they are always converted into /k/, /f/, and /s/ respectively.
- 4) Three non-graphemes ⟨-⟩, ⟨\*⟩, and ⟨space⟩ are encoded into the same binary codes since they have a same function in a word, i.e. no grapheme.

The modified partial orthogonal binary encoding for 26 graphemes and 3 non-graphemes (\*, -, and space) are listed in table II. Based on this encoding, two graphemes in a same small group have two different bits and the Euclidean distance

TABLE II: Modified Partial Orthogonal Binary Encoding for 26 Graphemes and 3 Non-graphemes

Grapheme	Modified partial orthogonal binary code
a	0011000
e	0010100
i	001001000
o	001000100
u	00100001000
b	110000001100
p	1100000010100
t	1100000000011000000000000000000000000000000000000000
d	11000000000101000000000000000000000000000000000000000
k	1100000000000011000000000000000000000000000000000000
q	1100000000000011000000000000000000000000000000000000
g	1100000000000010100000000000000000000000000000000000
c	11000000000000000110000000000000000000000000000000000
j	1100000000000000000101000000000000000000000000000000
f	1100000000000000000011000000000000000000000000000000
v	1100000000000000000011000000000000000000000000000000
s	1100000000000000000000001100000000000000000000000000
x	11000000000000000000000000000001100000000000000000000
z	11000000000000000000000000000001010000000000000000000
m	1100000000000000000000000011000000000000000000000000
n	11000000000000000000000000000101000000000000000000000
h	11000000000000000000000000000001110000000000000000000
r	11001100000000
l	11001100000000
w	11001100000
y	110001100
*	0100011
-	0100011
space	0100011

between them is  $\sqrt{2}$ . Those in different small group, but in a same main category, have four different bits and their Euclidean distance is 2. A vowel and a non-grapheme or a consonant and a non-grapheme have five different bits and their Euclidean distance is  $\sqrt{5}$ . A vowel and a consonant have six different bits and their Euclidean distance is  $\sqrt{6}$ . This encoding is expected to improve the capability of PNNR in classifying a pattern of grapheme sequence into a correct phoneme.

### C. Phonemic Rule-based Phoneme Filtering

The fifteen phonemic rules, to filter some potential phonemes to be selected by PNNR, are listed in table III. The first column is premise evaluating the focus grapheme (FG) and the first contextual graphemes on the left (L1) and right (R1). The second one is consequent containing one or more impossible phonemes (IP) to filter the possible conversions.

A grapheme is subject to two or more rules so that number of possible conversion to be lower. For example, grapheme  $ja\zeta_i$  can be converted into four different phonemes: /a/, /ai/, /aʊ/, or /a+ʔ/. But,  $ja\zeta_i$  in grapheme sequence  $jaku\zeta_i$  is subject to both rule 1 and 2 so that the grapheme is not possible to be converted into phoneme /ai/ nor /aʊ/. Hence,  $ja\zeta_i$  in the grapheme sequence  $jaku\zeta_i$  is only to be converted into /a/ or /a+ʔ/.

### D. Pseudo Nearest Neighbour Rule

The PNNR used here works in the same way as in [9]. It selects the best possible phonemes by finding the minimum total distance between the current pattern and the possible phonemes (classes) taking into account the  $k$  closest patterns. The total distance is calculated using equation 1, where  $u_j$  is the weight for the  $j$ -th neighbour,  $L$  is the contextual length,  $d_{li}$



TABLE III: Phonemic Rules to Filter Some Potential Phonemes (Conversion Classes), where FG is the Focus Grapheme, IP is Impossible Phoneme(s), L1 and R1 is the First Contextual Grapheme on the Left and the Right Respectively

Premise	Consequent
FG is ⟨a⟩ and R1 is not member {(i),(y)}	IP is /au/
FG is ⟨a⟩ and R1 is not member {(u),(w)}	IP is /au/
FG is ⟨e⟩ and R1 is not member {(i),(y)}	IP is /eu/
FG is ⟨e⟩ and R1 is not member {(a),(e),(i),(o),(u)}	IP is {/ε+ʔ/,/ε+ʔ/}
FG is ⟨g⟩ and L1 is not member {(n)}	IP is /*f/
FG is ⟨i⟩ and L1 is not member {(a),(e),(o)}	IP is /*f/
FG is ⟨i⟩ and R1 is not member {(a),(e),(o)}	IP is /i+ʔ/
FG is ⟨k⟩ and R1 is not member {(h)}	IP is /x/
FG is ⟨n⟩ and R1 is not member {(c),(j),(s)}	IP is /jʌ/
FG is ⟨n⟩ and R1 is not member {(g),(k)}	IP is /ŋ/
FG is ⟨o⟩ and R1 is not member {(i),(y)}	IP is /ou/
FG is ⟨s⟩ and R1 is not member {(y)}	IP is /ʃ/
FG is ⟨u⟩ and L1 is not member {(a)}	IP is /*f/
FG is ⟨u⟩ and R1 is not member {(a),(e),(o)}	IP is /u+ʔ/
FG is ⟨y⟩ and L1 is not member {(n),(s)}	IP is /*f/

and  $d_{ri}$  are the distances of the  $i$ -th contextual grapheme on the left and right calculated using the modified partial orthogonal binary grapheme encoding, and  $w_i$  is the weight for the  $i$ -th contextual grapheme.

$$T = \sum_{j=1}^k u_j \sum_{i=1}^{L/2} (d_{li}w_i + d_{ri}w_i) \quad (1)$$

The neighbourhood weight for the  $j$ -th neighbour,  $u_j$ , is formulated by equation 2, where  $c$  is an power constant as introduced in [9].

$$u_j = \frac{1}{j^c} \quad (2)$$

The graphemic contextual weight used here is an exponentially decaying function as proposed in [9]. It is formulated by equation 3, where  $w_i$  is the weight for the  $i$ -th contextual grapheme,  $p$  is an exponential constant, and  $L$  is the graphemic contextual length distributed equally into left and right of the focus. Thus, the first contextual grapheme has the maximum weight since it is the most important grapheme in deciding the best phoneme, whereas the last one has the minimum.

$$w_i = p^{L/2-i+1} \quad (3)$$

### III. EXPERIMENTAL RESULTS

The dataset used in this research is a set of 50K words with corresponding pronunciation (phonemic symbols) collected from the great dictionary of the Indonesian language (*Kamus Besar Bahasa Indonesia Pusat Bahasa* KBBI) third edition, released in 2008, developed by *Pusat Bahasa*. To get a valid evaluation, the PNNR-based G2P model is tested using 5-fold cross-validation. First, the dataset is randomly split into five subsets {1, 2, 3, 4, 5} of 10K different words each. Next, five datasets: A, B, C, D, and E, are developed to contain 40K for parameter tuning and 10K for evaluation. Dataset A consists of {1, 2, 3, 4} for parameter tuning and {5} for evaluation, dataset B consists of {1, 2, 3, 5} for parameter tuning and {4} for evaluation, and so on.

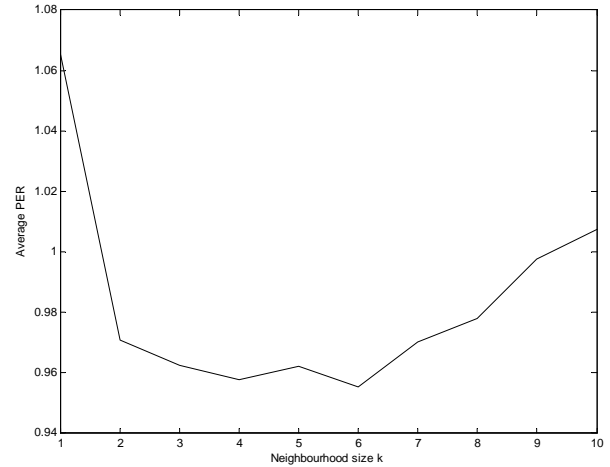


Fig. 2: Average PER of PNNR-based G2P for Varying Neighbourhood Size  $k$

#### A. Optimum Parameters

Three parameters of PNNR, the neighbourhood size  $k$ , the power constant of neighbourhood weight  $c$ , and the exponential constant for graphemic contextual weight  $p$ , are tuned using the five fold datasets. The first parameter to be tuned is  $k$  since it is so varying based on the problem that it is difficult to be predicted, while the optimum  $c$  is assumed to be 1.0 and the optimum  $p$  is predicted to be 2.0. Here the PNNR is evaluated for varying  $k$  with  $c = 1.0$  and  $p = 2.0$ . The result in figure 2 shows that when  $k = 1$  produces very high PER since considering only one neighbour can lead PNNR to be a too general classifier. It also gives high PER when considers so many neighbours that make it too specific in classifying a pattern. The PNNR produces the lowest PER on  $k = 6$ .

Next, the PNNR with  $k = 6$  and  $p = 2.0$  is analyzed using varying  $c$ . The result illustrated by figure 3 shows that when  $c$  is less than 1.0 the PNNR yields high PER since a low  $c$  makes the closest neighbour has quite similar distance to the further ones. It also produces high PER when  $c$  is 1.7 or more since a high  $c$  makes the closest neighbour has too high distance and the further ones are too low. It produces the lowest PER when  $c = 1.6$ .

The PNNR with  $k = 6$  and  $c = 1.6$  is then evaluated using varying  $p$ . The result in figure 4 shows that very small  $p$ , less than 1.7, make the PNNR yields high PER because the closest contextual graphemes has a similar importance to the further ones. It gives the lowest average PER, around 0.93%, when  $p = 1.9$ .

#### B. Modified Grapheme Encoding and Phonemic Rule

Next, the PNNR-based G2P using modified grapheme encoding and phonemic rule is cross-validated using five fold datasets. The experimental result in table IV shows that it significantly reduces both PER, up to 13.07%, and WER, around 12.13%, relative to the Indonesian PNNR-based G2P proposed in [9].

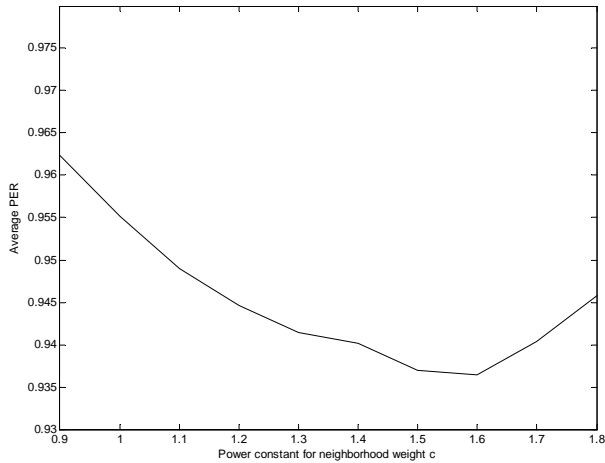


Fig. 3: Average PER of PNNR-based G2P for Varying Power Constant for Neighbourhood Weight  $c$

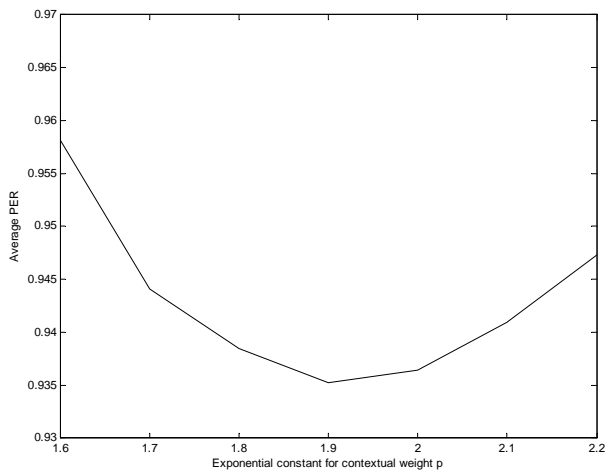


Fig. 4: Average PER of PNNR-based G2P for Varying Exponential Constant for Contextual Weight  $p$

TABLE IV: Advantage of Modified Grapheme Encoding and Phonemic Rule for PNNR-based G2P

PNNR-based G2P	PER (%)	WER (%)
Without phonemic rule	1.07	7.67
With phonemic rule and modified grapheme encoding	0.93	6.74

### C. Most Errors

As predicted in the introduction, the grapheme ⟨e⟩ produces most errors, up to 82%. This is caused by two problems: 1) Four prefixes, 'ber', 'me', 'per', and 'ter', produce many ambiguous conversions with basic words. For example, a grapheme ⟨e⟩ in a basic word 'berang' (irascible) is converted into /ɛ/, but ⟨e⟩ in a derivative word 'berangin' (windy) should be pronounced as /ə/ since 'ber' is a prefix for the basic word 'angin' (wind) that is always pronounced as /bɛr/. A

grapheme ⟨e⟩ in a basic word 'memang' (indeed) is converted into /ɛ/, but ⟨e⟩ in a derivative word 'memangsa' (to prey) should be pronounced as /ə/ since 'me' is a prefix for the basic word 'mangsa' (prey) that is always pronounced as /mə/. A grapheme ⟨e⟩ in a basic word 'peroksida' (peroxide) is converted into /ɛ/, but ⟨e⟩ in a derivative word 'perokok' (smoker) should be pronounced as /ə/ since 'pe' is a prefix for the basic word 'rokok' (cigarette) that is always pronounced as /pə/. A grapheme ⟨e⟩ in a basic word 'pering' (tuberculosis) is converted into /ɛ/, but ⟨e⟩ in a derivative word 'teringat' (remembered) should be pronounced as /ə/ since 'ter' is a prefix for the basic word 'ingat' (remember) that is always pronounced as /tɛr/; 2) A grapheme ⟨e⟩ in some similar compound words can be dynamically pronounced as either /ɛ/ or /ə/, such as a word 'termoelektris' (thermo-electric) that is pronounced as /tɛrmœɛtɛktris/, but 'termoelektrisitas' (thermo-electricity) should be pronounced as /tɛrmœɛtɛktrisitas/. A word 'reses' (recess) is pronounced as /rɛsɛs/, but 'resesi' (recession) is pronounced as /rɛsɛsi/. Such cases are very hard to be solved by the PNNR.

### IV. CONCLUSION

A 5-fold cross-validation shows the modified partial orthogonal binary grapheme encoding and phonemic rule are capable of reducing the relative PER by 13.07%. A detail analysis shows the most errors, up to 82%, come from grapheme ⟨e⟩ that can be dynamically converted into either /ɛ/ or /ə/ since four prefixes, 'ber', 'me', 'per', and 'ter', produce many ambiguous conversions with basic words and also from some similar compound words with both different pronunciations for grapheme ⟨e⟩, such as a word 'reses' (recess) that is pronounced as /rɛsɛs/ but 'resesi' (recession) is pronounced as /rɛsɛsi/. In the future, a stemming procedure can be incorporated to find the basic words from the derivatives with the four prefixes in order to reduce such errors.

### ACKNOWLEDGMENT

The first author is an employee of Telkom Foundation of Education (Yayasan Pendidikan Telkom, YPT) as a lecturer at School of Computing, Telkom University (former: Telkom Institute of Technology). Now he is pursuing a doctoral program on Computer Science, Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Gadjah Mada University. This work is supported by YPT with grant number: 15/SDM-06/YPT/2013.

### REFERENCES

- [1] A. V. D. Bosch and W. Daelemans, "Data-oriented methods for grapheme-to-phoneme conversion," in *The sixth conference on European chapter of the Association for Computational Linguistics (EACL)*. Morristown, NJ, USA: Association for Computational Linguistics, 1993, pp. 45–53. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=976744.976751>
- [2] F. Yvon, P. B. D. Mareu, C. Alessandro, M. Bagein, G. Bailly, F. Be, S. Foukia, V. Auberger, J. Goldman, E. Keller, D. O. Shaughnessy, V. Pagel, B. Zellner, F. Sannier, and J. Ve, "Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French," *Computer Speech & Language*, vol. 12, no. 4, pp. 393–410, 1998.
- [3] P. Taylor, "Hidden Markov Models for Grapheme to Phoneme Conversion," in *INTERSPEECH*, 2005, pp. 1973–1976.

- [4] U. D. Reichel and F. Schiel, "Using Morphology and Phoneme History to improve Grapheme-to-Phoneme Conversion," in *INTERSPEECH*. ISCA, 2005, pp. 1937–1940.
- [5] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.
- [6] D. Wang and S. King, "Letter-to-sound Pronunciation Prediction Using Conditional Random Fields," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 122–125, 2011.
- [7] R. Rasipuram, M. M. Doss, and L. Epfl, "Combining Acoustic Data Driven G2P and Letter-to-Sound Rules for Under Resource Lexicon Generation," in *INTERSPEECH*. ISCA, 2012, pp. 1818–1821.
- [8] A. Hartoyo and Suyanto, "An improved Indonesian grapheme-to-phoneme conversion using statistic and linguistic information," *International Journal Research in Computing Science (IJRCS)*, vol. 46, no. 1, pp. 179–190, 2010.
- [9] Suyanto and A. Harjoko, "Nearest neighbour-based Indonesian G2P conversion," *Telkonnika (Telecommunication, Computing, Electronics, and Control)*, vol. 12, no. 2, pp. 389–396, 2014.
- [10] H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A. M. Moeliono, *Tata Bahasa Baku Bahasa Indonesia (The Standard Indonesian Grammar)*, 3rd ed. Jakarta: Balai Pustaka, 1998.
- [11] A. Chaer, *Fonologi Bahasa Indonesia (Indonesian Phonology)*. Jakarta: Rineka Cipta, 2009.

# Testing and Analysis of Activities of Daily Living Data with Machine Learning Algorithms

Ayse Cufoglu and Adem Coskun  
Faculty of Science and Technology  
Department of Engineering  
University of Westminster  
London, UK

**Abstract**—It is estimated that 28% of European Union's population will be aged 65 or older by 2060. Europe is getting older and this has a high impact on the estimated cost to be spent for older people. This is because, compared to the younger generation, older people are more at risk to have/face cognitive impairment, frailty and social exclusion, which could have negative effects on their lives as well as the economy of the European Union. The 'active and independent ageing' concept aims to support older people to live active and independent life in their preferred location and this goal can be fully achieved by understanding the older people (i.e their needs, abilities, preferences, difficulties they are facing during the day). One of the most reliable resources for such information is the Activities of Daily Living (ADL), which gives essential information about people's lives. Understanding this kind of information is an important step towards providing the right support, facilities and care for the older population. In the literature, there is a lack of study that evaluates the performance of Machine Learning algorithms towards understanding the ADL data. This work aims to test and analyze the performance of the well known Machine Learning algorithms with ADL data.

**Keywords**—Activities of Daily Living (ADL); Machine Learning (ML); Classification Algorithms; Active and Independent Aging

## I. INTRODUCTION

Europe is becoming older as a result of lower birth rates and higher life expectancy. According to Eurostat [1], older people will form 28% of the European Union's (EU's) population by 2060 and people that are aged 80 and over will rise from 5% to 12% which is very close to the percentage (around 15%) that is estimated for young people (aged 0-14) in 2060. It is known that older population is more vulnerable towards cognitive impairment, frailty and social exclusion, which could negatively effect the health-care system as well as older people's Quality of Live (QoL) and independence. This reality triggered many investments that are focused on the areas that promote and facilitate 'active and independent ageing', which aims to provide the needed support (i.e. medical support, monitoring services) to let older people live active and independent life in their preferred locations. Activities of Daily Living (ADL) is one of the important parameters that gives essential information about people's lives. Understanding ADL information of older people is an important step forward understanding their:

- needs and abilities,
- difficulties they are facing during their everyday lives

- detecting new medical conditions to provide the right support and personalized care for them.

This work aims to test and analyze the performance of the well known Machine Learning (ML) algorithms with ADL data. This study will provide an opportunity to understand how ML algorithms cope with the ADL data with respect to learning, analyzing, understanding and classifying the ADL activities. Following ML algorithms have been considered for this study: Instance Based Learner (IBL), K Nearest Neighbour (KNN or IBK), KStar (K\*), J48, Locally Weighted Learning (LWL) and Naive Bayesian Tree (NBTree).

This paper is organized as follows; Section II provides a literature review for this work. Section III provides details about the ML algorithms that are tested and analyzed within this study. Section IV presents an information about the datasets that have been used for the simulations. Information about the evaluation measures are presented in Section V. Section VI presents the simulation results and their evaluations. Finally, Section VII provides conclusion and future works for this study. Throughout this paper records and instances are used as synonyms.

## II. BACKGROUND

Activities of Daily Living (ADL) defines the activities we perform in our daily living, such as self care (i.e. feeding ourselves, bathing, dressing, grooming), work and leisure activities [2]. Statistics about the EU's older people population for the coming years influenced many works within the health-care domain. There have been various works in the literature focused at the ADL information (i.e. ADL information capture and analysis). For instance, work by Cheng et al. [3] proposed an ADL recognition engine, called the Adaptive Learning Hidden Markov Model (ALHMM), based on Hidden Markov Model combining the Viterbi and Baum-Welch algorithms for enhanced accuracy and learning capabilities. According to the authors, their proposed model provides an effective and practical solution for ADL recognition. The ALHMM's inference efficiency outperforms the Hidden Markov Model (HMM). In another work [2], Fleury et al. classify the data collected through various sensors, such as temperature, hygrometry and infrared presence sensors, into one of the ADL (hygiene, toilet use, eating, resting, sleeping, communication, and dressing/undressing) using Support Vector Machines (SVM) classifier. The ADLs of 13 young and healthy volunteers have

been recorded and used in this study. Results presented in this work showed that the SVM with polynomial kernel achieved 75% accuracy while SVM with Gaussian kernel achieved 86%. In another work Tapia et al. [4] proposed an alternative sensing system to detect activities in real homes. Different from the traditional sensing technology, the proposed system uses simple, low cost “tape on and forget” sensors to recognize individual’s activities. In this study, the authors used Naive Bayesian (NB) algorithm for activity detection. Here, NB extended to incorporate temporal relationships among sensor firings and recognize activities. The data that has been used for this study collected from two subjects within two home (one-bedroom apartments) settings. Here, one home was fitted with 77 sensors and the other with 84 sensors. The experimental testing with small dataset showed that the detection accuracies ranged from 25% to 89% depending on the evaluation criteria used.

As it can be seen from the aforementioned works, there is a lack of work in the literature that compares different ML algorithms with respect to their performance with ADL data. Considering this gap in the literature, this paper focuses on the testing, analysis and comparison of the well-known ML algorithms’ performance on ADL data.

### III. MACHINE LEARNING ALGORITHMS

For this work, performance of IBL, KNN, KStar (K\*), LWL, J48 and NBTree classification algorithms have been analysed with the ADL dataset. Following paragraphs give information about each of these algorithms.

#### A. Lazy Classifiers

1) *Instance Based Learner (IBL)*: The Nearest Neighbour (NN) classifier uses normalized Euclidean distance to compare each test instance with the training instances. The closest training instance predicted to has the same class label with the test instance [5] [6] [7]. In case of more than one training instance qualified as the closest, the class label of the first one is assigned to be the class label of the test instance [8]. The IBL is a comprehensive form of NN. In IBL the comparison between the test instance and the training instance is done as follows;

Assume training instance  $X_i = \{x_i(1), x_i(2), \dots, x_i(B)\}$  and test instance  $Y_j = \{y_j(1), y_j(2), \dots, y_j(B)\}$ . Here comparison between training instance and test instance is done feature by feature as;

- If feature is numeric,

$$g(y_j(b), x_i(b)) = (y_j(b) - x_i(b))^2 \quad (1)$$

- If feature is symbolic,

$$g(y_j(b) - x_i(b)) = \begin{cases} 0, & \text{if } y_j(b) = x_i(b) \\ 1, & \text{if } y_j(b) \neq x_i(b) \end{cases} \quad (2)$$

where the  $g(y_j(b), x_i(b))$  function shows the similarity between  $b$  values of the training and test instances and the distance is calculated as;

$$dist(Y_j, X_i) = \sqrt{\sum_{b=1}^B g(y_j(b), x_i(b))} \quad (3)$$

2) *K Nearest Neighbor (KNN)*: K Nearest Neighbor (KNN), also known as IBK, is one of the well known IBL algorithm, where different from IBL,  $K$  closest instances are retrieved and the label of the majority class among these instances is assigned as the class label for the test instance [5] [6] [9]. The following paragraphs are associated with IBK:

- If the class attribute is symbolic then the class label of the test instance  $Y_j$  is same as the class label of the highest vote among the  $K$  nearest neighbours. For a scenario, where  $K = 3$ , if the three nearest neighbours  $X_1, X_2$  and  $X_3$  belong to the classes,  $C_1, C_1$  and  $C_2$  respectively, then  $C_1$  is assigned as the class label for  $Y_j, Y_j \in C_1$ , since  $C_1$  is the predominant class label among nearest neighbours.
- If class attribute is numeric then the class label of the test instance will be the mean of the nearest neighbours. Following the above assumptions, label of  $Y_j, L(Y_j)$ , is calculated as;

$$L(Y_j) = \frac{\sum_{m=1}^K L(X_m)}{K} \quad (4)$$

where  $L(X_m)$  represents the class label of  $X_m$ .

3) *KStar (K\*)*: K\* is based on entropy distance measure where the distance between two instances is defined as the complexity of transforming one instance into another [6] [10]. This complexity calculation is done in two steps. First step is to define the finite set of transformations that map instances to instances. A ‘program’ which transforms one instance  $Y_j$  to another instance  $X_i$  is a finite sequence of transformations starting at  $Y_j$  and terminating at  $X_i$  [10]. Kolmogorov is one of the well known entropy distance measures where the distance between two instances is the shortest string connecting them [10]. Hence, this approach is focused on the shortest transformation out of many possible transformations. Here, the resulted distance measure is very sensitive to small changes. For this problem K\* is defined as the distance of summing all possible transformations between two instances.

4) *Locally Weighted Learning*: Locally Weighted Learning (LWL) is a weighted IBL that assigns weights to instances using the IBL and uses these locally weighted training instances for classification [11] [12]. While IBK performs local approximation for each test instance  $Y_j$ , LWL performs explicit approximation of  $L(Y_j)$  for region surrounding  $Y_j$  by fitting linear function and quadratic to K nearest neighbours.

#### B. Decision Trees

Decision Trees (DTs) are data structures that can examine the data and induce a classification tree and its rules to make predictions [13]. A successful classification with the DTs requires well-defined classes and pre-classified training data [14]. The classification accuracy on the training data set and the size of the tree affect the quality of the DT. Construction of the tree model incorporates into two-phases; building phase and pruning phase. The building phase includes a series of divisions of the training dataset that is carried out based on the decision rules [14]. This partitioning is continued until the resulted classes have homogenous instances. In the pruning phase, on the other hand, the nodes that may cause over fitting and low accuracy are pruned [6] [14].

1) *Naive Bayesian Tree (NBTree)*: Naive Bayesian (NB) classifier is one of the Bayesian classifier algorithms. In many works it has been proven that NB classifiers are one of the most computationally efficient and simple algorithms for ML and DM applications [15]-[19]. Unlike BN, NB classifiers assume that all attributes within the same class are independent, given the class label. Based on this assumption, which also reduces the computational complexity of BN classifier, the NB classifier modifies the Bayesian rule as follows [16] [20]:

- Assume  $Cl_b$  as class label where  $Cl_b = \{Cl_1, Cl_2, \dots, Cl_B\}$  and  $b = 1, 2, \dots, B$
- Assume  $Y_j$  as unclassified test instance where  $Y_j = \{y_j(1), y_j(2), \dots, y_j(A)\}$  for  $k = 1, 2, \dots, A$

Based on these assumptions the NB is calculated as;

$$P(Cl_b|Y_j) = \arg \max_{Cl_b} P(Cl_b)P(y_j(1), y_j(2), \dots, y_j(A)|Cl_b) \quad (5)$$

$$P(Cl_b|Y_j) = \arg \max_{Cl_b} \prod_{k=1}^A P(y_j(k)|Cl_b) \quad (6)$$

NBTree classifier generates a DT with NB classifier at the leaves [6].

2) *J48*: J48 is another DT algorithm. The J48 classification algorithm is the enhanced version of C4.5 decision tree and has been developed to generate a pruned or un-pruned C4.5 DT [6] [21]. The C4.5 DT uses divide-and-conquer approach to growing DTs. The J48 has been developed to address the problems of both C4.5 and Id3 [8] classifiers.

#### IV. DATASETS

For this study we have used the ADL dataset provided by Bruno et al. [22] at the UCI Machine Learning Repository [23]. This dataset includes accelerometer data for Human Primitive detection and it is a collection of labeled accelerometer data recordings that was collected through the tri-axial wrist-worn accelerometer. The data transmission from accelerometer to the PC is done via wired USB cable where the dataset is a collection of 14 ADLs (i.e. brussteeth, climbstairs, combhair) performed by the 16 volunteers (11 Male and 5 female) aged between 19 and 81. For this study, 11 ADLs of the three volunteers (2 male and 1 female) have been chosen from the complete dataset. The chosen ADLs for our study are brussteeth, climbstairs, combhair, drinkglass, getupbed, liedownbed, pourwater, sitdownchair, standupchair, usetelephone and walk. For the simulations we have created three datasets from the complete dataset. The first dataset uses only nine datapoints (x,y,z parameters) to represent each performed activity. This dataset has 8495 total records with 11 attributes. Table I shows example records from this dataset. In the second dataset, each performed activity represented with 15 datapoints. This dataset has 5091 records and each record has 17 attributes. Table II shows few records from dataset2 and Fig.1 shows the count of each activity in dataset2. The third dataset, on the other hand, uses 30 datapoints to represent each performed activity. This dataset has 2538 records and 32 features to represent each record.

TABLE I: Two sample records from ADL dataset1 with three datapoints

x1	y1	z1	x2	y2	z2	x3	y3	z3	gender/id	HMP
22	49	35	22	49	35	22	52	35	f1	brussteeth
29	44	40	35	11	28	41	13	24	m1	combhair

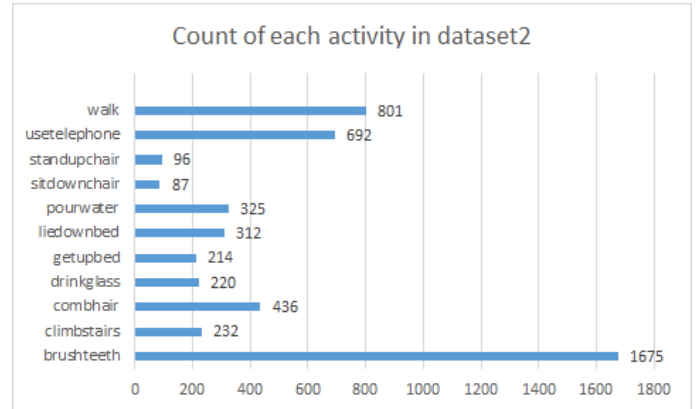


Fig. 1: Count of each ADL activities within the dataset2

#### V. EVALUATION MEASURES

There are a number of evaluation measures that are commonly used to evaluate the performance of each classification algorithm. Here, we focused on six evaluation measures that are True Positive Rate (TPR), False Positive Rate (FPR), Precision (or Positive Predictive Value (PPV)), Recall, F-measure (F1) and Receiver Operating Characteristics (ROC) curve. These evaluations are mainly done based on four parameters that are True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). Simplest way to explain each of these parameters is through an example.

Given an instance and the classifier, there can be four outcomes [24];

- If an instance is positive and it is predicted to belong to a positive class then this is counted as TP,
- If a positive instance predicted to belong to a negative class then it is counted as FN,
- If an instance is negative and it is predicted to belong to a negative class then this is counted as TN, but,
- If the same instance is predicted to belong to a positive class then this is counted as FP.

Based on these parameters each of the evaluation measures is calculated as follows [24][25];

$$\begin{aligned} TPR(\text{or Recall}) &= TP/(TP + TN) \\ FPR &= FP/(FP + TN) \\ Precision &= TP/(TP + FP) \\ F1 &= 2TP/(2TP + FP + FN) \end{aligned} \quad (7)$$

Following this the accuracy is calculated as;

$$Accuracy = (TP + TN)/(P + N) \quad (8)$$

TABLE II: Two sample records from ADL dataset2 with fifteen datapoints

x1	y1	z1	x2	y2	z2	x3	y3	z3	x4	y4	z4	x5	y5	z5	gender/id	HMP
34	48	49	35	48	45	34	51	45	34	48	47	35	49	45	m2	pourwater
27	46	46	26	43	46	26	48	46	26	48	47	28	48	45	m1	drinkglass

TABLE III: Classification accuracy results of the algorithms with nine datapoints

Algorithm	Correctly classified instances (%)
IBL	69.05
Kstar	70.86
KNN	68.65
LWL	52.61
NBTree	64.60
J48	65.96

where (P+N) corresponds to the sum of positive and negative instances (the total number of instances). Here, each of the measure can have a value between 0 and 1. For TP, Precision, Recall, F-measure and ROC, the high accuracy means values closer to 1. For FP, as the accuracy gets higher the value of FP gets closer to 0. Precision is one of the performance measures and differs from accuracy as it does not relate to the true value (accepted reference value) [26]. Precision shows the closeness of the independent test results on homogeneous data and usually computed as a standard deviation of the results [26]. The ROC curves are two dimensional graphs where Y axis represents TPR and X axis represents FPR and the graph depicts the relative tradeoffs between the TP (benefits) and FP (costs) [24].

## VI. SIMULATIONS

For the simulations, datasets that are mentioned in Section IV have been used. The 10-fold cross-validation has been chosen as the test mode and for each simulation the same dataset has been used as a training and test datasets. Here, 'HMP' is chosen as a class attribute for all simulations. Each of the six ML algorithms are trained on the same training set and tested on the same test set (i.e for the first simulation all algorithms trained and tested on dataset1).

Table III shows the accuracy performance of all six algorithms with dataset1. Here it can be seen that the lazy classifiers have mixed accuracy results where some are quite high and some are quite low. DTs, on the other hand, was consistent on performing average classification accuracies. From the results, it can be seen that the best result is achieved by the Kstar with 70.86% accuracy (6020 instances classified correctly). This result is followed by the IBL, which classified 5866 instances correctly and 2629 instances incorrectly. The KNN algorithm performed very similar results with IBL by classifying 68.65% of the instances correctly. DT algorithms followed the results of lazy classifiers by classifying 5604 (J48) and 5372 (NBTree) instances correctly. The lowest accuracy is achieved by the LWL algorithm. Here LWL classified 52.61% of the instances correctly. Hence, from the first simulation results it can be said that the highest and lowest results both achieved by the lazy classifiers.

Second simulations has been carried out with dataset2.

TABLE IV: Classification accuracy results of the algorithms with fifteen datapoints

Algorithm	Correctly classified instances (%)
IBL	69.75
Kstar	70.53
KNN	69.57
LWL	53.13
NBTree	64.60
J48	64.74

Table IV shows the results of these simulations. From the results it can be seen that the DT based algorithms achieved around 64% accuracy where NBTree classified 3289 (64.60%) instances correctly and J48 classified 3296 (64.74%) instances correctly. In general, lazy classifiers performed with higher accuracy results compared to the DT based algorithms. According to the results, the best classification result is archived by the Kstar algorithm with 70.53% of instances classified correctly. This results is followed by the IBL algorithm with %69.75 correctly classified instances. The KNN classifier closely followed the result of IBL by classifying 3542 (69.57%) out of 5091 instances correctly. Within all six algorithms, the lowest classification accuracy is archived by LWL. Here LWL classifier classified 2705 instances correctly and 2386 instances incorrectly, which corresponds to 46.86% of the total instances. We carried out additional number of simulations only with the KNN algorithm to see if changing the number of neighbors (N) affects the classification performance of the KNN algorithm with ADL dataset2. Fig.2 shows the results of these simulations. From the figure it can be seen that the classification accuracy increases slightly between N=2 (69.57%) and N=4 (71.83%). After that the accuracy stays around 71% between N=4 and N=8. For N=9 and N=10 the classification accuracy of the KNN drops again and stays around 70%. In general, results showed that increasing the number of neighbours improved the accuracy of KNN but very little. Hence, it can be said that increasing the number of neighbours did not have a great positive impact, with respect to the accuracy performance, on the KNN. Moreover, by comparing the KNN results with IBL, it can be said that using more than two nearest neighbours for classification achieved slightly (%2.08 when N=4) better results than using only one neighbour (N=1).

Table V shows the time taken for each algorithm to build the classification model for the second simulations. The time taken to build the model is the system time that was used to run the classifier and is converted from millisecond into seconds by WEKA. These results shows that, in general, the lazy classifiers took less time to build the classification model than DT classifiers. Here, the NBTree algorithm took the longest time, 20.95, compared to all other five algorithms. The NBTree algorithm is a hybrid algorithm which carries both DT and NB algorithms' working principles, and therefore it was expected from it to require more time. The lowest time requirement (0



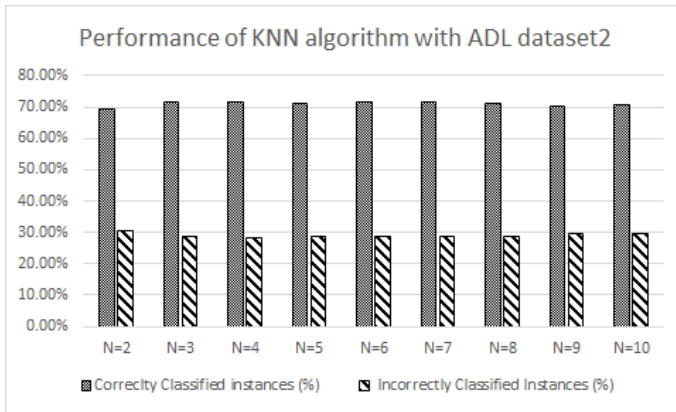


Fig. 2: Performance of KNN algorithm with ADL dataset2

TABLE V: Algorithms vs. time taken to build the model with fifteen datapoints

Algorithm	Time
IBL	0.02 sec
Kstar	0 sec
KNN	0 sec
LWL	0 sec
NBTree	20.95 sec
J48	0.52 sec

TABLE VI: Detailed accuracy measures of the algorithms' performance with fifteen data points

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
IBL	0.698	0.043	0.696	0.698	0.695	0.827
Kstar	0.705	0.044	0.7	0.705	0.7	0.948
KNN	0.696	0.046	0.709	0.696	0.693	0.872
LWL	0.531	0.17	0.342	0.531	0.411	0.859
NBTree	0.646	0.062	0.641	0.646	0.629	0.898
J48	0.647	0.054	0.641	0.647	0.644	0.823

sec), on the other hand, is archived by the KNN (N=2), Kstar and LWL. Within lazy classifiers, only IBL needed little bit more time than 0 sec. The J48 needed 0.52 sec to build the classification model, which is quite less than what NBTree required.

Table VI shows the detailed evaluation measure results of all six algorithms with dataset2. From this table it can be seen that accuracy results of all six algorithms are reflected and justified with the evaluation measures too. Kstar, which has the highest classification accuracy results with the second dataset, has the TP, Precision, Recall and F-measure values around 0.7 and ROC area around 0.9, all of which are the closest to 1 among all other algorithms. Moreover, LWL, which performed the lowest classification accuracy, has the values for the same evaluation measures between 0.3 and 0.5 that are the ones closest to 0 among all other algorithms.

Final simulations have been carried out with the third dataset that has 30 datapoints that corresponds to 10 sets of x,y and z compositions. As previously mentioned, this dataset have 32 features to define each instance where 30 out of 32 features are datapoints. Table VII shows the simulation results

TABLE VII: Classification accuracy results of the algorithms with thirty datapoints

Algorithm	Correctly classified instances (%)
IBL	69.70
Kstar	68.71
KNN	68.71
LWL	54.60
NBTree	64.38
J48	61.74

TABLE VIII: Detailed accuracy measures of the algorithms' performance with thirty data points

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
IBL	0.697	0.042	0.7	0.697	0.696	0.827
Kstar	0.687	0.043	0.694	0.687	0.687	0.946
KNN	0.695	0.046	0.708	0.695	0.691	0.871
LWL	0.546	0.143	0.367	0.546	0.437	0.861
NBTree	0.644	0.06	0.634	0.644	0.624	0.882
J48	0.617	0.06	0.613	0.617	0.615	0.802

for all six algorithms. From the table it can be seen that the best results, among all six algorithms, is achieved by the IBL algorithm with 69.70% classification accuracy. This result very closely followed by the KNN algorithm with 69.50% accuracy which corresponds to 1769 correctly classified instances and 774 incorrectly classified instances. KStar, which achieved the best classification results in first (9 datapoints) and second (15 datapoints) simulations, classified 68.71% (1744) instances correctly with third dataset. Approximately with 4% less instances, NBTree classified 64.38% instances correctly. Similar to the first two simulations, LWL performed the lowest classification accuracy by classifying only 54.60% (1386) of the total instances correctly. From the results it can be seen that, different from the previous simulations with dataset1 and dataset 2, the best classification accuracy results are achieved by the IBL algorithm rather than Kstart and this has been reflected on the evaluation measure results that are presented on Table VIII. Here, IBL has the values for TP, Precision, Recall, F-measure and ROC area closest to 1 compared to all other algorithms. Moreover, the FP value for the IBL is the closest to 1 when compared with all other five algorithms.

## VII. CONCLUSION AND FUTURE WORKS

Europe is aging and understanding the elderly people plays very critical role in finding the best solutions for demographic challenges. The ADL information of the elderly can give insight into elderly people's lives with respect to what they do during the day and what kind of problems they encounter, such as when they start showing symptoms for the new medical conditions. This information then could be used to formulate the best personalized solutions which will make it possible for them to live independent, health and active life. As the popularity of wearable devices is in rapid increase, such devices could be used to gather ADL information from the elderly people. However, analyzing and understanding such information is as important as the collection of the information. In this paper we tested, analyzed and compared the performance of well-known lazy and decision tree classification algorithms with the ADL data. The simulation results with three different sized ADL

datasets showed that, in general, lazy classifiers performed better than DT algorithms with respect to understanding and classifying the ADL data more accurately with very small time requirements. The results also showed that the ML algorithms could be used for such purposes. On the other hand, the maximum classification accuracy result, reflected through several evaluation measures, was around only 71%, which is still not good enough considering we are dealing with very sensitive data. Hence, as a future work we would focus on investigating different ways of improving the accuracy performance of the algorithms which resulted the best performances in this study.

## REFERENCES

- [1] The 2015 Aging Report, Underlying assumptions and projection Methodologies, European Commission (DG ECFIN) and Economic Policy Committee (AWG), ISSN 0379-0991 (print), ISSN 1725-3217 (online), 2014, Available [online]: <http://tinyurl.com/o7dnhhs>
- [2] A. Fleury, M. Vacher and Norbert Noury. SVM-based multimodal classification of activities of daily living in health smart homes:sensors, algorithms and first experimental results, 14(2), pp. 274-283, March 2010.
- [3] B. C. Cheng, Y. A. Tsai, G. T. Liao and E. S. Byeon. HMM machine learning and inference for Activities of Daily Living recognition,The Journal of Supercomputing, Springer, 54(1), pp. 29-42, October 2010.
- [4] E. M. Tapia, S. S. Intille and K. Larson. Activity recognition in the home using simple and ubiquitous sensors, Proceedings of Pervasive, vol. 3001, Springer-Verlag, Berlin/Heidelberg, pp. 158-175, April 2004.
- [5] D. W. Aha, D. Kibler and M. K. Albert. Instance-based learning algorithms, Machine Learning Journal, 1(6), pp. 37-66, January 1991.
- [6] I. H. Witten, E. Frank and M. A. Hall. Data mining practical machine learning tools and techniques 3rd Edition, Morgan Kaufmann, USA, pp. 472-550, 2011.
- [7] A. Cufoglu, M. Lohi and C. Everiss. Weighted Instance Based Learner (WIBL) for user profiling, IEEE Symposium on Applied Machine Intelligence and Informatics, pp. 201-205, January 2012.
- [8] I. H. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
- [9] O. Gomez, E. F. Morales and J. A. Gonzales. Weighted instance-based learning using representative intervals, Mexican International Conference on Advances in Artificial Intelligence, vol. 4827, pp. 420-430, November 2007.
- [10] J. G. Clear and L. E. Trigg. K\*: An instance-based learner using an entropic distance measure, International Conference on Machine Learning, pp. 108-114, July 1995.
- [11] C.G. Atkeson, A.W. Moore, and S. Schaal. Locally weighted learning. Artificial Intelligence Review, 11(1):11-73, 1997.
- [12] P. Englert. Locally Weighted Learning, Seminar Class on Autonomous Systems, 2012.
- [13] V. P. Bresferean. Analysis and predictions on students behaviour using decision trees in WEKA environment, IEEE International Conference on Information Technology Interfaces, pp. 51-56, June 2007.
- [14] A.L. Symeonidis and P.A. Mitkas. Agent intelligence through data mining multiagent systems, artificial societies and simulated organizations; 14, Electron. Book, New York: Springer Science and Business Media, pp. 21-23, 27-28, 2005.
- [15] D. R. Amancio, C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. A. Rodrigues and L. D. F. Costa. A Systematic Comparison of Supervised Classifiers. PLoS ONE 9(4), April 2014.
- [16] Z. Shi, Y. Huang and S. Zhang. Fisher score based naive Bayesian classifier, IEEE International Conference on Neural Networks and Brain, pp. 1616-1621, October 2005.
- [17] J. Huang and C. X. Ling. Using AUC and accuracy in evaluating learning algorithms, IEEE Transactions on Knowledge and Data Engineering, 17(3), pp. 299-310, March 2005.
- [18] Z. Xie and Q. Zhang. A study of selective neighbourhood-based naive Bayes for efficient lazy learning, IEEE International Conference on Tools with Artificial Intelligence, pp. 758-760, November 2004.
- [19] G. Santafe, J.A. Loranzo and P. Larranaga. Bayesian model averaging of naive bayes for clustering, IEEE Transactions on Systems, Man, and Cybernetics, 36(5), pp. 1149 -1161, October 2006.
- [20] F.V. Jensen. Introduction to Bayesian networks, Springer, August 2000.
- [21] J. R. Quinlan. C4.5: Programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
- [22] B. Bruno, F. Mastrogiovanni and A. Sgorbissa. ADL recognition with wrist-worn accelerometer dataset, February 2014. Available [online]:<http://archive.ics.uci.edu/>
- [23] M. Lichman. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, 2013. Available[online]:<http://archive.ics.uci.edu/ml>
- [24] T. Fawcett. An introduction to ROC analysis, Pattern Recognition Letters, Elsevier, 27(8), pp. 861-874, June 2006.
- [25] D. M. W. Powers. Evaluation:From precision,recall and F-factor to ROC, informedness, markedness & Correlation, Journal of Machine Learning Technologies, 2(1), pp. 37-63, December 2011.
- [26] M. Ester, H.P. Kriegel, J. Sander and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise, International Conference on Knowledge Discovery and Data Mining, pp. 226-231, August 1996.

# A Context-Aware Recommender System for Personalized Places in Mobile Applications

Soha A.El-Moemen Mohamed  
Demonstrator

Information System Department,  
Faculty of Computer and Information  
Assiut University, Egypt

Taysir Hassan A.Soliman  
Associate Professor

Information System Department,  
Faculty of Computer and Information  
Assiut University, Egypt

Adel A.Sewisy  
Professor

Computer Science Department,  
Faculty of Computer and Information  
Assiut University, Egypt

**Abstract**—Selecting the most appropriate places under different context is important contribution nowadays for people who visit places for the first time. The aim of the work in this paper is to make a context-aware recommender system, which recommends places to users based on the current weather, the time of the day, and the user's mood. This Context-aware Recommender System will determine the current weather and time of the day in a user's location. Then, it gets places that are appropriate to context state in the user's location. Also, Recommender system takes the current user's mood and then selects the location that the user should go. Places are recommended based on what other users have visited in the similar context conditions. Recommender system puts rates for each place in each context for each user. The place's rates are calculated by The Genetic algorithm, based on Gamma function. Finally, mobile application was implemented in the context-aware recommender system.

**Keywords**—recommender system, context, context-aware, genetic algorithm, gamma function.

## I. INTRODUCTION

The massive scale of data on the Internet today makes it difficult for users to find relevant information. Thus, the information is customized according to the users' needs and preferences. Often the application of recommender systems uses Collaborative filtering and content of the list. In a scenario of mobile devices, customization of information is more important, because of the restrictions on mobile devices about the displays and input capabilities, bandwidth, etc. For example, the traveler with PDA, smart phones or pocket PCs needs access to the weather at the destination or have a recommendation for a hotel in the neighboring region. It is recommended to customize not only the use of user profiles in advance, but also a context such as the current location, time of the day, the current weather, or the mood of the user. Context can be defined as the information that can be used to describe the situation entities i.e., person, place, or subject, which is relevant to the interaction between the user and the application, including the user and the application themselves [8]. The first challenge to generate context recommendations is learned via identifying the contextual factors such as weather that affect the ratings and thus that are worth considering. Second, get a representative set in the context of Rates ,i.e., rankings under different contextual circumstances is free from the context rankings. Finally, the expansion of conventional recommender systems is to exploit

the additional information fields in the evaluation context. In this paper, a novel algorithm was performed , called Context-Aware Genetic Recommender System (CAGRS). CAGRS system consists of five modules: 1-Collecting Stay Points, 2-Clustering Stay Points, 3-Determining Popular Places in each cluster, 4- Getting the weather and time of the day in each Stay Point, and 5- Getting frequency of the weather and time of the day in each cluster. CAGRS take user's parameters, such as (current user's location, user's mood) and determine the current weather and time of the day. CAGRS uses Genetic Algorithm for predicting rates of each place based on gamma function as the fitness function. Genetic Algorithm (GA) considered as the optimization algorithm which gives an optimal solution to most problems. In the field of Artificial Intelligence, GA is a search heuristic that uses the process of natural selection (survival of the fittest) [9]. Genetic algorithm puts rates in each place according to different context conditions and for each user. The contributions of this paper are:

- 1- Developing a context-aware recommender system based on a Genetic algorithm with gamma function.
- 2- Rating places by using number of visits for many users to each place and by examining context conditions in each one.
- 3- Recommending the best place that user should go based on the user's mood or the current weather and time of the day in user's location.

This paper is divided as follows: Section 2 contains summarization of related work; Section 3 contains steps of how the system runs; Section 4 discusses experimental results and discussions. Finally, Section 5 concludes the paper and presents future works.

## II. RELATED WORK

Yong Zheng and Bamshad Mobasher [19] proposed context match as a new contextual modeling technique and study different ways to represent context similarity and include it into the recommendation. They showed how context similarity can be incorporated into the sparse linear method and matrix factorization algorithms.

Qiang Liu and ShuWu [6] proposed Contextual Operating Tensor (COT) Model, which denoted the public semantic effects of contexts as a contextual operating tensor and denotes a context as a hidden vector. Then, to model the semantic operation of a context grouping, they created contextual operating matrix from the contextual operating tensor and hidden vectors of contexts. Thus, hidden vectors of users and items can be used by the contextual operating matrices.

Wolfgang Woerndl and Christian Schueller [12] recommended mobile applications to users based on what other users have used in the same context. The idea is to make a hybrid recommender system to manage the extra complexity of context. Users can choose among different content-based or collaborative filtering factors, including a rule-based module, using information on point-of-interests in the vicinity of the user, and a factor for the mixing traditional collaborative filtering.

Sabri Boutemedjet and Djemel Ziou [2] proposed a new framework for the recommendation of context-aware visual documents by modeling the needs of users, and context, as well as the collection of visual document together in a unified model. It also directed the user need for a variety of recommendations.

Matthias Braunhofer, Mehdi Elahi, and Francesco Ricci [3] presented a new context-aware mobile recommender system for places of interest (POIs). Unlike current systems, which study users' preferences solely from their earlier ratings, it also considers their personality - using the Five Factor Model. Personality developed by asking users to complete a short and entertaining questionnaire as part of the registration the process, and then used in: (1) an active learning element actively acquires ratings-in-context for POIs that users are likely to have practiced, hence decreasing the stress and annoyance to rate (or skip rating) items that the users don't know, and (2) in the recommendation model that builds based on matrix factorization and, therefore, can deal with unrated items.

Keith Cheverst and Nigel Davies[13] defined their skills of developing and estimating GUIDE, an intelligent electronic tourist guide. The GUIDE system has built to overwhelmed many of the limits of the old information and navigation tools available to city visitors. For example, group-based tours are inherently inflexible with fixed starting times and static durations and are constrained by the need to satisfy the interests of the benefits rather than the specific interests of individuals. Following a period of needs capture, involving experts in the field of tourism, they developed and installed a System for use by visitors to Lancaster. The system combines mobile computing technologies with a wireless infrastructure to present city visitors with information tailored to both their personal and environmental contexts.

Barry Brown and Matthew Chalmers [4] described as a light weight portable system designed for the exchange of entertainment. This system allowed Visitors to the city to share experiences with others, through disc Computers that share pictures, audio, and location. Collaborative filtering algorithm used historical data in previous visits that use images, the web pages and recommended places for visitors, bringing the media together online with the city's streets.

Pedro G. Campos and Ignacio Fernandez-Tobias [5] conducted an empirical comparison of several pre-filtering, post-filtering and contextual modeling approaches on the movie recommen-

dation field. To acquire confident contextual information, they made a user questionnaire where participants were asked to rate movies, stating the time and social companion with which they preferred to watch the rated movies.

Chein-Chung Hwang<sup>1</sup>, Yi-Ching Su [11] described a new recommender system, which performed a genetic algorithm to learn personal preferences of customers and provide tailored suggestions.

Karthik Srinivasa Gopalan, Senthil Nathan developed the use of context capture on the users devices as a method of learning all the user activity patterns and using these patterns to generate content recommendations. They proposed a new [9]recommendation system based on an evolutionary algorithm that evaluates new content based on multiple objectives.

Daniel Siewiorek, Asim Smailagic [1] developed SenSay which is a context-aware mobile phone that adjusts to dynamically changing environmental and physiological shapes. In addition to operating ringer volume, vibration, and phone alerts, SenSay can provide faraway callers with the ability to communicate the urgency of their calls, make call offers to users when they are idle, and provide the caller with reaction on the current status of the SenSay user. Some sensors including accelerometers, light, and microphones are mounted at numerous points on the body to provide data about the users context.

Mark van Setten and Stanislav Pokraev [15] defined the context-aware mobile tourist application COMPASS that adjusts its services to the users requirements based on both the users interests and his current context. To provide context-aware recommendations, a recommender system has been joined by a context-aware application stage. They described how this integration has accomplished and how users felt about such an adaptive tourist application.

Anand Ranganathan and Roy H. Campbell [14] argued that computing environments must offer middle ware support for context-awareness. They also planned a middle ware that facilitates the progress of context-aware agents. The middle ware allows agents to gain contextual information easily, reason about it using different logics and then adjust themselves to changing contexts. Another key issue in these environments is letting autonomous, varied agents to have a common semantic understanding of contextual information. This middle ware attacks this problem by using ontologies to define different kinds of contextual information.

Jakob E. Bardram [17] presented the strategy of a context-aware pill container and a context-aware hospital bed, both of which responds and adapts according to what is happening in their context. The applications have been assessed in a number of workshop with clinicians and patients. Based on this experiential work of designing, developing, and evaluating context-aware clinical applications, the paper plans some key design principles for a context-awareness framework, supporting the development and deployment of context-aware clinical computer applications.

Yong Jae Lee and Kristen Grauman [18] proposed to power information about earlier learned categories to enable more correct discovery. They introduced a new object graph descriptor to encode the layout of object-level co-occurrence designs about an unfamiliar area, and showed that by using it to model the communication between image's known and unknown objects they can better notice new visual categories. Rather than mine for all categories from scratch, their method

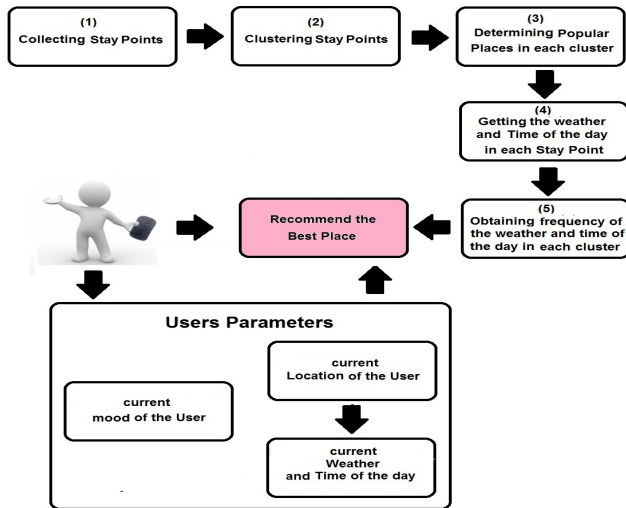


Fig. 1: System Model

can frequently identify new objects while drawing on useful signs from familiar ones.

Xiping Hu, Xiong Li [10] proposed a multidimensional context-aware social network design which aims to offer a mobile ecosystem to enable context awareness in the development and utilization of mobile crowd sensing applications. This mobile ecosystem is made to provide context awareness skills for different roles in the system and facilitate interactions between them. This system can ease the expansion of context-aware mobile applications, and enable context-aware mobile crowd sensing considering environmental, personal, and social information. They offered a flow of context-aware results designed on this system, and highlight the orchestrations and advantages of different context-aware structures in the system for different types of users in mobile crowd sensing.

Po-Han Wu and Gwo-Jen Hwang [20] provided a context-aware mobile learning system established for nursing training courses. During the learning events, each student is equipped with a mobile device; moreover, sensing devices are used to sense whether the student has conducted the operations on the accurate location of the dummy patients body for evaluating the physical status of the specified disease. The learning system not only guides specific students to make each operation of the physical valuation procedure on dummy patients, but also provides direct feedback and extra materials to them if the operations or the operating sequence is incorrect.

### III. CONTEXT-AWARE GENETIC RECOMMENDER SYSTEM (CAGRS)

In this section, CAGRS system will be discussed in details and illustrate all system's modules as shown in figure 1. CAGRS algorithm consists of five main modules: 1-Collecting Stay Points, 2-Clustering Stay Points, 3-Determining Popular Places in each cluster, 4- Getting the weather and time of the day in each Stay Point, and 5- Getting frequency of the weather and time of the day in each cluster. These modules will be explained in details as follows:

#### A. CAGRS Modules

1) *Collecting Stay Points*: From the work stated in [16], the stay points have resulted from GPS trajectories that are collected by Microsoft Research Asia's Geolife project<sup>1</sup>. To calculate latitude and longitude of a stay point, the average will use as GPS points that are around specific locations, as illustrated for the following equations (1) and (2):

$$SP_{lat} = (\sum_{r=x}^y p_r.lat) / (|p|) \quad (1)$$

$$SP_{long} = (\sum_{r=x}^y p_r.long) / (|p|) \quad (2)$$

$SP_{lat}$  and  $SP_{long}$  are average latitude and longitude of GPS points in stay point in sub-trajectories. These stay locations might be hotels, restaurants, historic places, homes, schools, etc. The total number of resulted stay points is 14391 for all users.

2) *Clustering Stay Points*: After collecting stay points, now these stay points need to be collected in different clusters. Clustering is performed using DBSCAN algorithm. Density-based Spatial Clustering of Applications with Noise (DBSCAN) is a density based clustering algorithm, using top-down approach. It groups together points that are packed together (points with many nearby neighbors). We choose DBSCAN algorithm because it does not need to know the number of clusters in the data as opposed to k-means. DBSCAN has a concept of noise, and is to deal with outliers. DBSCAN needs just two parameters and is mostly insensitive to the ordering of the points in the database. By clustering stay points using DBSCAN algorithm, points that are close to each other with distance less than  $\epsilon$  will be obtained in the same cluster and others considered as noise points. All 14391 stay points were collected in 172 clusters if  $\epsilon=30$  meters [9] as shown in figure 2.

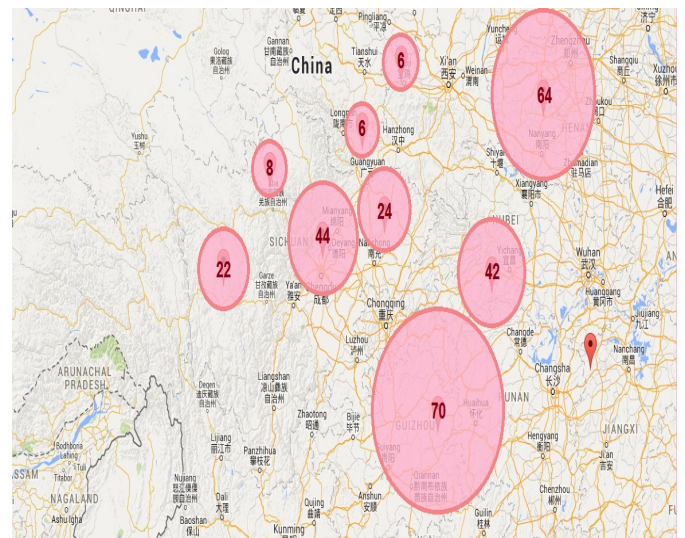


Fig. 2: Stay Point Clusters

<sup>1</sup><http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/>



3) *Determining the Popular Places in each cluster:*

Now, the center of each cluster needed to be determined. Equations (3) and (4) are used to calculate the latitude and longitude for centers by calculating the mean value of all stay points(latitude, longitude) in each cluster.

$$C_i.lat = (\sum_j sp_j.lat)/(|sp_i|) \quad (3)$$

$$C_i.long = (\sum_j sp_j.long)/(|sp_i|) \quad (4)$$

Now, every point in the center of each group must be converted into meaningful titles by getting the names of each place. The reverse geocoding process takes latitude and longitude as input and returns the site of this point as output. There are different APIs like Google Geocoding API, My Geoposition API, GeoNames API, Nominatim open stream map, etc. GeoNames API was used for mapping each point in its address <sup>2</sup>. Can not set all points in the name of the place. Only 166 locations from 172 points marked with their places names.

4) *Getting the weather and time of the day for each stay point in the cluster:* In this step, the history of weather needs to be determined and the time of the day in each stay point (location). There are many APIs to determine this information (i.e., Open Weather Map, World Weather Online, National Climatic Data Center,etc.). All these APIs take the location information as input (latitude, and longitude) and take the date information (time, date) and get all weather information as output. This paper used (Weather Underground API) <sup>3</sup>.

5) *The Optioning frequency for the weather and time of the day in clusters:* Finally, in this step calculates the most frequent weather and time of the day in each cluster. Each popular place belongs to each cluster will be marked by the most frequent weather and time of the day in this cluster.

TABLE I: Weather and time of day in each location

Place	Weather	Time of the Day
Gulf of Masira, gulf,Oman	Storm	Night
Nubrongcarog, farm village, Tibet Autonomous Region,China	Mostly Cloud	Morning
Songqinma, grazing area, Tibet Autonomous Region,China	Clear	Night
Cuochoae, grazing area, Tibet Autonomous Region,China	Partly,Cloud	Morning

B. *User Parameters*

As shown in figure 1, CAGRS system takes user parameters, such as (Weather, time of the day and user's mood). As an example, if the mood of the user is (Sad), then the user may need to go to cinema to watch a comedy movie. The system ask user to put his/her current mood, then it recommends the best appropriate place for a user's mood. for example, if the weather is sunny the user may need to go to a place with air condition, if the weather is rainy, the user need closed place. As shown in Table I, each popular place stored with weather and time of the day.

<sup>2</sup><http://ws5.geonames.org>

<sup>3</sup><http://www.wunderground.com>

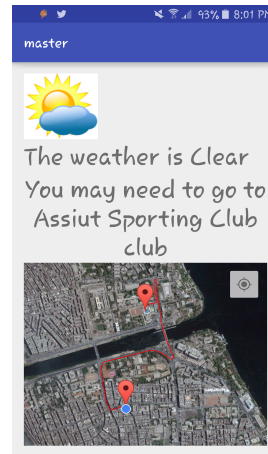


Fig. 3: Recommended Place based on the current weather

TABLE II: Appropriate places for each user's mood

Place	Mood
Gulf of Masira, gulf,Oman	Happy
Nubrongcarog, farm village, Tibet Autonomous Region,China	Angry
Naren, grazingarea, Tibet Autonomous Region,China	Excited
Surbhi, hotel, Madhya Pradesh, India	Tired

To get this information in to practice, the location of the user needs to be determined and the current weather/time of the day in a user's location using Google map API. When the current Weather/Time of the day in a user's location match the Weather/Time of the day for a popular place, CAGRS will recommend this popular place to this user and will show the direction between user's location and the recommended popular place, as shown in Figure 3. As shown in Table II, each type of user's mood stored with the most appropriate place. Figure 4 shows how to take the mood of the user and show the recommended place according to this mood.

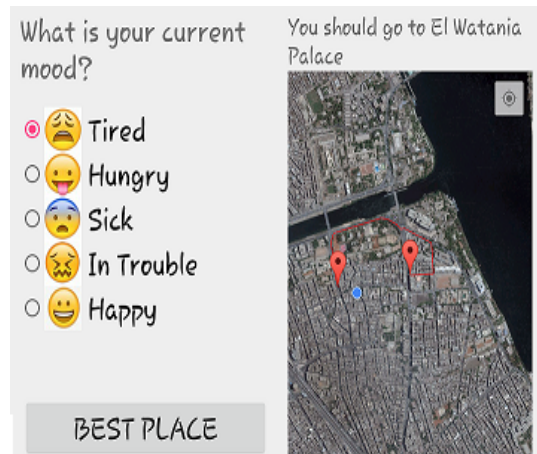


Fig. 4: Recommended Place based on the current user's mood

C. Context-aware Recommender System

Finally, Recommender System will give rates for each place according to each user and each context state. There are many methods used to determine these rates (i.e., Genetic Algorithm, Matrix Factorization). This paper used Genetic Algorithm [9]. The resulted data stored in 3D Cube with the following dimensions (user, place, context) as shown in Figure 5. Figure 6 shows the Place Rating Matrix that present rates for each place based on users and context. The value of the place's rate is calculated using the frequency of each context state (weather, time of the day, and user's mood) in each cluster. For unknown place rating values, The Genetic algorithm will predict these values.

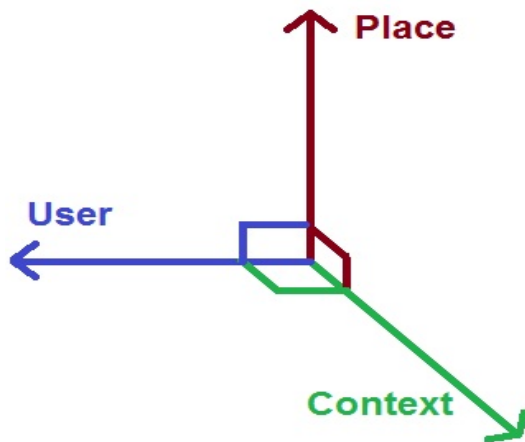


Fig. 5: Recommender System dimensions

Place	User	Context	Rate
P <sub>1</sub>	U <sub>1</sub>	C <sub>1</sub>	R <sub>1</sub>
P <sub>2</sub>	U <sub>2</sub>	C <sub>2</sub>	?
...	...	...	...
P <sub>n</sub>	U <sub>n</sub>	C <sub>n</sub>	R <sub>n</sub>

Fig. 6: Place Rating Matrix

The Genetic algorithms (GAs) are stochastic search methods that guide a population of solutions using the principles of growth and natural genetics. Extensive research has been achieved exploiting the robust properties of genetic algorithms and representing their capabilities across a broad spectrum of optimization problems, including feature selection and weighting tasks. GAs are modeled loosely on the principles of the evolution via natural selection, employing a population of individuals that undergo selection in the

TABLE III: Experimental Results

ID	User	Place	Context	Rate
1	1	2	storm	13
2	2	2	cloudy	47
3	3	12	sunny	187
4	4	2	clear	116

presence of variation-inducing operators such as mutation and crossover. A fitness function is used to evaluate individuals, and generative success varies with fitness. Architecture of Genetic Algorithm is as follows:

1-Initial population:

Two Random Place Rating matrices will generate from  $R_{min}$  to  $R_{max}$ . This initial population's format was chosen based on Place Rating Matrix's format illustrated in figure 6.

2-Fitness Function:

Fitness function used in genetic algorithm is a Gamma Function as in equation (6).

$$\Gamma(x) = \int_0^{\infty} s^{x-1} e^{-s} ds \quad (6)$$

3-Crossover method:

Offspring is generated from two matrices parents by using uniform crossover. Offspring is created from selected individuals by selecting random place  $P_i$ , random user  $U_j$  and random context  $C_k$ . Then, the block of first parent matrix that before i, j, and k is exchanged with block of second parent matrix that after i, j, and k.

4-Selection approach:

Selection of the best individuals is based on the minimum value of fitness function. Algorithm will be stopped when the value of fitness function reaches the minimum value for specific matrix. The resulted matrix is called optimal Place Rating Matrix with no null values.

IV. COMPARISONS AND EXPERIMENTAL RESULTS

In this paper, the Genetic Algorithm was performed that based on Gamma Function on 166 places, 182 users, and 12 different context states. The results as shown in Table III, are the final matrix that consists of (User, Place, Context, Rate). The results consists of 19116 records. Experimental results are run on Intel Core i5 Processor, 4.00 GB Ram and Windows 7 operating system. PHP programming language and MATLAB were used for obtaining the results. We divided the dataset in 5 small datasets with 3823 records in each dataset. Table IV and Figure 7 show the comparisons when performing 1 run ,3 runs,5 runs and 10 runs in each 5 small datasets. To compare the results with data from users' reviews, 1539 responses were collected by making a simple questionnaire. Figure 8 shows the frequency of places in each weather state.

We measured the Mean Absolute Error (MAE) as in equation 7 between the resulted data from algorithm and the real data that was taken by questionnaire.

$$MAE = \frac{1}{n} \sum_{i=1}^n |e_i| \quad (7)$$

Table IV shows the rating value of each weather state



TABLE IV: MAE comparing results

Place	Real Data						Experimental Data						MAE
	Clear	Hazy	Mostly Cloudy	Partly Cloudy	Sunny	Storms	Clear	Hazy	Mostly Clouds	Partly Clouds	Sunny	Storms	
Home	103	121	113	112	112	117	488	100	198	362	456	434	233.6666667
open area	104	100	102	103	103	100	300	200	100	150	400	211	125.5
farm village	103	101	102	101	101	102	372	100	725	774	347	124	305.6666667
halting place	101	100	100	102	102	100	300	322	443	212	121	221	169
grazing area	101	104	101	101	101	103	433	132	121	221	257	211	127.3333333
hill	104	100	100	101	101	100	100	100	100	100	106	100	1.666666667
hotal	100	111	117	110	110	117	121	302	233	238	213	244	114.3333333
Island	103	101	100	101	101	100	323	123	156	129	321	130	96
lake	102	100	102	102	102	100	431	239	178	165	212	133	125
mountain	102	100	100	101	101	100	433	122	246	138	111	103	91.5
oil field	101	101	100	101	101	100	100	100	100	100	122	121	7.5
peak	105	100	100	101	101	100	102	100	102	100	102	100	1.166666667
stream	103	100	101	101	101	100	233	323	415	332	123	322	190.3333333
swamp	101	100	100	101	101	100	126	439	443	412	147	187	191.8333333
well	106	100	101	101	101	100	125	256	321	157	323	221	132.3333333

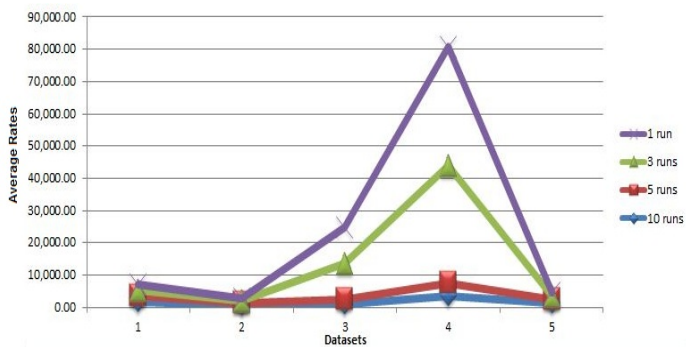


Fig. 7: 1 run, 3 runs, 5 runs and 10 runs in 5 datasets

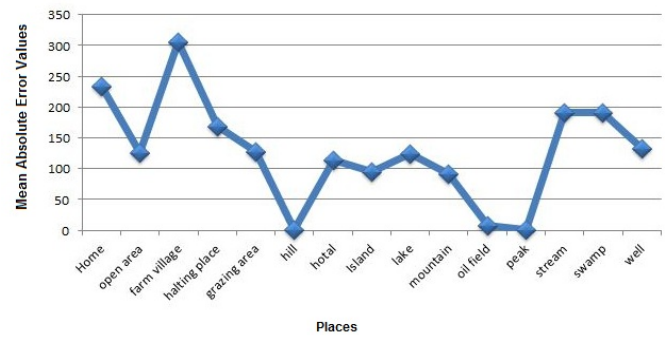


Fig. 9: MAE results

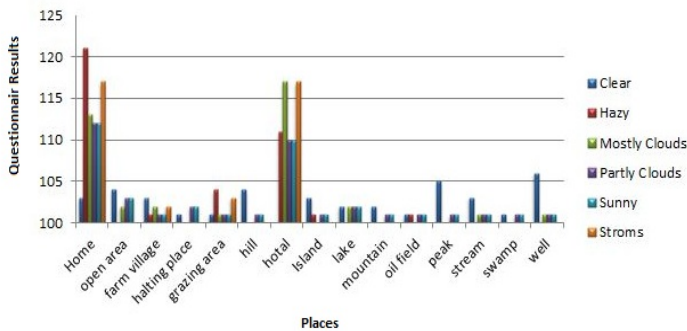


Fig. 8: Real data results

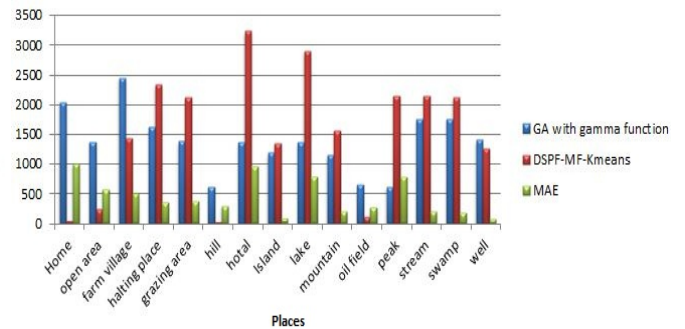


Fig. 10: MAE results

for each place from real data results and from experimental data results and shows the MAE value between the two results. Figure 9 shows the MAE results between on Google questionnaire results and experimental from CAGRS algorithm results.

We compared results generated from Genetic Algorithm (GA) based on Gamma function with the resulted data from Distributional-Semantics Pre-filtering (DSPF) based on Matrix Factorization (MF) and K-Means clustering algorithm [7]. Figure 10 and Table V show the MAE between two sets.

We performed in CAGRS system a Gamma Function as fitness function in Genetic Algorithm. In Table VI and Figure 11, The results that based on gamma function compared

with results from Root Mean Square error in [9] and were compared with Prediction Accuracy in [14].

## V. CONCLUSION AND FUTURE WORK

In this paper, the context aware mobile recommender system learns user preferences from the past ratings as well as their personality. This information is utilized to generate high-quality recommendations for places of interest in light of contextual goal situation. The implementation of the system have described, regarding design, recommendation logic, user interface, and features. for future work, it is nice to make many improvements to the system. First, recommendations to the user need to be made when the current situation seems

TABLE V: the MAE between GA with gamma function and DSPF-MF-Kmeans

Place	GA with gamma function	DSPF-MF-Kmeans	MAE
Home	2038	34	1002
open area	1361	231	565
farm village	2442	1433	504.5
halting place	1619	2332	356.5
grazing area	1375	2111	368
hill	606	21	292.5
hotel	1351	3233	941
Island	1182	1343	80.5
lake	1358	2901	771.5
mountain	1153	1554	200.5
oil field	643	102	270.5
peak	606	2143	768.5
stream	1748	2144	198
swamp	1754	2114	180
well	1403	1254	74.5

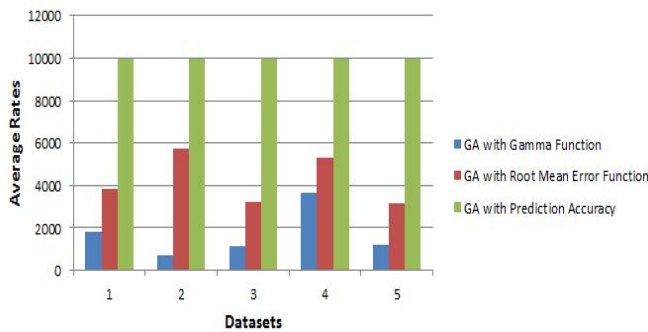


Fig. 11: Comparing results among gamma function, root mean square error, and prediction accuracy

TABLE VI: Comparing results among gamma function, root mean square error, and prediction accuracy

Datasets	GA with gamma function	GA with Root Mean Square Error	GA with Prediction Accuracy
1	1819.352676	3854.154606	9961.9474
2	726.6724821	5730.233999	9961.8807
3	1129.015292	3224.995895	9962.2678
4	3666.113198	5325.618697	9961.8812
5	1188.293084	3159.17148	9962.3953

appropriate, without relying on explicit user's request. In addition, taking advantage of human emotions in the process of recommendation and know the current user activity that can be drawn from wearable devices such as smart watches and smart chips. Finally, developing roads to determine the user's personal conclusion by the personal characteristics of the users of the features of Facebook or using a mobile phone.

#### REFERENCES

[1] Bardram, J. E. (2004). Applications of context-aware computing in hospital work: examples and design principles. Proceedings of the 2004 ACM symposium on Applied computing, ACM.

[2] Boutemedjet, S. and D. Ziou (2008). "A graphical model for context-aware visual content recommendation." Multimedia, IEEE Transactions on 10(1): 52-62.

[3] Braunhofer, M., et al. "STS: A Context-Aware Mobile Recommender System for Places of Interest."

[4] Brown, B., et al. (2005). Sharing the square: collaborative leisure in the city streets. ECSCW 2005, Springer.

[5] Campos, P. G., et al. (2013). Context-aware movie recommendations: An empirical comparison of pre-filtering, post-filtering and contextual modeling approaches. E-Commerce and Web Technologies, Springer: 137-149.

[6] Cheverst, K., et al. (2000). Developing a context-aware electronic tourist guide: some issues and experiences. Proceedings of the SIGCHI conference on Human Factors in Computing Systems, ACM.

[7] Codina, V., et al. (2015). "Distributional semantic pre-filtering in context-aware recommender systems." User Modeling and User-Adapted Interaction: 1-32.

[8] Dey, A. K., et al. (2001). "A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications." Human-computer interaction 16(2): 97-166.

[9] Gopalan, K. S., et al. (2011). A context aware personalized media recommendation system: an adaptive evolutionary algorithm approach. Bio-Inspired Computing: Theories and Applications (BIC-TA), 2011 Sixth International Conference on, IEEE.

[10] Hu, X., et al. (2014). "Multidimensional context-aware social network architecture for mobile crowdsensing." Communications Magazine, IEEE 52(6): 78-87.

[11] Hwang, C.-S., et al. (2010). Using genetic algorithms for personalized recommendation. Computational Collective Intelligence. Technologies and Applications, Springer: 104-112.

[12] Lee, Y. J. and K. Grauman (2012). "Object-graphs for context-aware visual category discovery." Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(2): 346-358.

[13] Liu, Q., et al. (2015). COT: Contextual Operating Tensor for Context-aware Recommender Systems. Twenty-Ninth AAAI Conference on Artificial Intelligence.

[14] Ranganathan, A. and R. H. Campbell (2003). A middleware for context-aware agents in ubiquitous computing environments. Middleware 2003, Springer.

[15] Siewiorek, D., et al. (2003). Sensay: A context-aware mobile phone. null, IEEE.

[16] Soliman, T. H. A., et al. (2015). Developing a mobile location-based collaborative Recommender System for GIS applications. Computer Engineering and Systems (ICCES), 2015 Tenth International Conference on, IEEE.

[17] Van Setten, M., et al. (2004). Context-aware recommendations in the mobile tourist application COMPASS. Adaptive hypermedia and adaptive web-based systems, Springer.

[18] Woerndl, W., et al. (2007). A hybrid recommender system for context-aware recommendations of mobile applications. Data Engineering Workshop, 2007 IEEE 23rd International Conference on, IEEE.

[19] Wu, P.-H., et al. (2012). "A Context-Aware Mobile Learning System for Supporting Cognitive Apprenticeships in Nursing Skills Training." Educational Technology and Society 15(1): 223-236.

[20] Zheng, Y., et al. (2015). Similarity-based context-aware recommendation. Web Information Systems Engineering WISE 2015, Springer: 431-447.

# Performance Enhancement of Patch-based Descriptors for Image Copy Detection

Junaid Baber, Maheen Bakhtyar, Waheed Noor, Abdul Basit, Ihsan Ullah  
Department of Computer Science and Information Technology  
University of Balochistan

**Abstract**—Images have become main sources for the information, learning, and entertainment, but due to the advancement and progress in multimedia technologies, millions of images are shared on Internet daily which can be easily duplicated and redistributed. Distribution of these duplicated and transformed images cause a lot of problems and challenges such as piracy, redundancy, and content-based image indexing and retrieval. To address these problems, copy detection system based on local features are widely used. Initially, keypoints are detected and represented by some robust descriptors. The descriptors are computed over the affine patches around the keypoints, these patches should be repeatable under photometric and geometric transformations. However, there exist two main challenges with patch based descriptors, (1) the affine patch over the keypoint can produce similar descriptors under entirely different scene or the context which causes “ambiguity”, and (2) the descriptors are not enough “distinctive” under image noise. Due to these limitations, the copy detection systems suffer in performance. We present a framework that makes descriptor more distinguishable and robust by influencing them with the texture and gradients in vicinity. The experimental evaluation on keypoints matching and image copy detection under severe transformations shows the effectiveness of the proposed framework.

**Keywords**—Content-based image copy detection, SIFT, CSLBP, robust descriptors, patch based descriptors

## I. INTRODUCTION

With the availability of Internet and powerful open access image editing tools, it has become so easy and convenient to edit the images and distribute to others. Millions of images are shared and uploaded daily on image sharing sites such as Facebook, Flickr, and ImageShake. Image databases are increasing exponentially in size which causes so many problems for efficient image indexing and retrieval such as image piracy and redundancy. Much of work has been reported to address these problems. However, partial image copy detection and near duplicate detection are still very challenging [1], [2].

Generally, Image retrieval applications can be categorized into three types: near duplicate image detection, image copy detection, and similar image detection. In near duplicate detection, the task is to detect all image copies along with many other challenges such as the same scene captured by different viewpoint or captured on the different time. Image copy detection and partial duplicate detection are interchangeably used [1] and this is the subset of near duplicate image detection — original image is altered with changes in color, scale, partial occlusion, rotation, etc. Finally, similar image detection tends to find those images which are similar based on their visual

contents, texture, or attributes. Similar image detection applications are widely used by many image search engines such as Google image. In these applications global features such as color histograms, variance, and image entropy are mostly used. In this paper, we mainly focus on content-based image copy detection which comprises of image copy detection and near duplicate image detection, these applications are potentially needed for piracy, copyright violation, and efficient image searching.

There are two famous techniques to prevent copyright violation; Watermarking and Content Based Copy Detection (CBCD), respectively. Watermarking technique involves embedding the information in the digital signal in such a way that it is difficult to remove [3] and carried along with the signal whenever distributed or shared. The information in watermark can be visible or invisible, whereas invisible information is widely used. There are many schemes proposed for digital watermarking such as spectrum watermarks [4], quantization watermarks [5], and blind detection watermarks [6]. Watermarking is the process of the identification of codes that store the information of the owner. The main challenges for watermarks are the translation, scale change, rotation, and cropping. For watermarking systems, signatures are extracted before distribution which makes this technique not feasible for on-line applications where digital signals are already distributed. Whereas, CBCD is the complementary approach to watermarking. In CBCD systems, robust signatures or fingerprints from query images are extracted, and then compared to the signatures or fingerprints of gallery images (database) to determine whether the query image has the copy in given gallery or not. One of the main advantages of CBCD system over watermarking is that signature extraction is not required before distribution, signature or fingerprint can be called the feature vector of the digital content. The feature vector should be distinguishable, robust and efficient [7]. The main challenge for CBCD systems is that the copy of digital content can be modified or transformed to deceive the copy detection systems. The list of top 10 challenge transformations are shown in Table I.

Local keypoint descriptors are widely used to represent the images in CBCD. Firstly, keypoints are detected from the images, these keypoints should have high repeatability under various affine and geometrical transformations. Some of the most successful algorithms for keypoint detection include SIFT [8], SURF [9], and Harris and Hessian affine

TABLE I  
LIST OF CHALLENGES FOR COPY DETECTION

#	Description
C1	Camcording
C2	Picture in picture
C3	Insertions of pattern
C4	JPEG compression
C5	Change of illumination
C6	Cropping
C7	Blurring
C8	Image flipping
C9	Text insertion
C10	Decrease in image quality – This includes combination of all nine challenges along with image shifting, contrast, and image morphing

keypoints [10]. Secondly, local patches around the keypoint are estimated [10]. Thirdly, robust and discriminative descriptors which are invariant to many transformations such as scale, rotation, affine distortion, 3D viewpoint change, image noise, and illumination change are computed [8]. Finally, image is represented by the set of local keypoint descriptors which are later used for image retrieval, object recognition, copy detection, augmented reality and many others.

The two main limitation for this practice are: (1) keypoint descriptor is the representation of local image patch which is centered to keypoint, this patch can produce the similar descriptor under entirely different scene or context (ambiguity), and (2) the similar patch can produce different descriptors under image noise (indiscrimination). The example of ambiguity and indiscrimination are shown Figure 1.

To overcome these limitation, geometric relationship between local features is widely used by researchers in the application of visual object categorization and copy detection [11]–[14]. It has been argued that the performance of local features is significantly improved from *bag of visual words* to *bag of pair of visual words* [12]. But there are some limitations such as (1) the combination of possible visual word pairs grow quadratically w.r.t the vocabulary size. To overcome this problem, different features are used on these pair of visual words [15]. (2) Different features can be treated as its second limitation as they require additional information like class labels and does not necessarily lead to better performance [15]. These two limitations are discussed in Morioka et al [14], where they propose a reverse technique. Instead paring the visual words, they pair the raw descriptors before learning the visual words. It has been experimentally shown that paring before visual words learning is more effective then visual word paring themselves [13], [14].

In this paper, we extend our previous work [16]. We propose a framework to improve the performance of patch based descriptors for CBCD. In our framework, the descriptors are enriched by geometrically spatial relationship around the keypoint region which makes descriptor more discriminative. The rest of the paper is organized as follows. In Section II, we briefly discuss some related work for content-based image copy detection. In Section III, two famous descriptors are explained, and different approaches to make them robust are

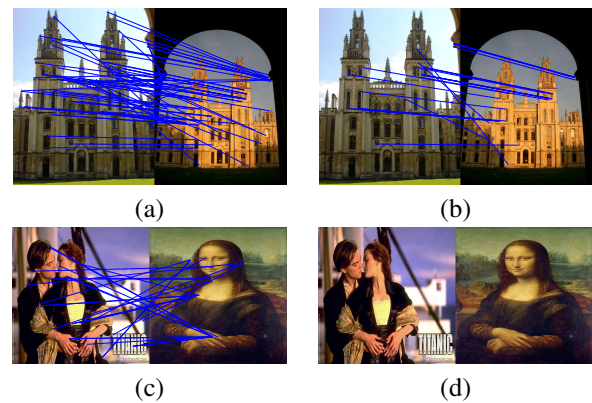


Fig. 1. SIFT descriptor matching, (a) and (c) SIFT descriptor, (b) and (d) neighbor base SIFT descriptor with two neighbors. Where (a) shows the indiscriminateness of SIFT descriptors and (c) shows the ambiguity.

discussed. In Section IV, detail experimental evaluation is given, and paper is concluded in Section V.

## II. RELATED WORK

This section is divided into two parts. In first part, we briefly discuss about image copy detection, and in second part, we briefly explain two famous patch based descriptors.

### A. Content-based image copy detection

Content-based copy detection has been boosted for last six year due to local features. Local features have been proven to be more resistant and robust for severe image transformations compared to global features. The descriptor SIFT [8] is also one of the major reasons for local features popularity. Many CBCD and image retrieval systems have been proposed based on SIFT and other local features such as GLOH [17], CSLBP [18], SURF [9], and BIG-OH [19].

Chang et al. [20] proposed RIME (Replicated IMage dE-tector) to detect pirated copies of images on Internet using wavelets and color space. The system has good accuracy for basic types of transformations. Kim [21] use Discrete Cosine Transform (DCT) for CBCD, as DCT is more robust to many distortions and changes in images. They converted the images into YUV format and only Y component is used in proposed method, as they argue that colors do not play important role in copy detection but colors are vital part in image retrieval (images similar based on color, texture, or objects). They successfully detected the copies of the test images with and without modifications, however they fail to detect the copies with 90° or 270° rotation [22]. Basit et al. [23], [24] proposed a method joint localization to track the target with unmanned vehicles. The method fuses the robot kinematics and target dynamics in single space model to produce better results. The global features are efficient for simple types of transformations, however, in case of severe transformations the performance of global features is very poor, for example, in case of cropping, occlusion, and aspect ratio change.

Xu et al. [25] proposed CBCD system based on SIFT and spatial features. They detect homogeneous and larger



circular patches using SIFT detector, and then multi-resolution histograms are computed as feature vectors. However, the performance of this system is poor in occlusion types of transformations. Zhou et al. [2] proposed a framework for partial image copy detection for large scale applications by using bag-of-visual-words model. They quantized the SIFT in descriptor space and orientation space. They encode the spatial layout of keypoints by XMAP and YMAP strategy, which helps to remove the outliers. However, their framework is not effective for object retrieval type of applications.

### B. Patch based descriptors

Most of the successful feature point descriptors described in the literature can be classified into two types: those based on gradient histograms [8], [9], [17], [26] and those based on local pixel intensity differences without explicit gradient calculations [18], [27]–[29]. We will call the latter group “intensity-based” descriptors. Two famous descriptors are used in the proposed framework for experiments: Scale-invariant feature transform (SIFT) [8], and Center-Symmetric Local Binary Pattern (CSLBP) [18].

1) *Center-Symmetric Local Binary Pattern*: CSLBP is an extension of Local Binary Patterns (LBP). In LBP, each pixel value  $p$  is compared with its  $N$  neighbors with radial distance  $R$ . If the pixel value of  $p$  is less than its neighbors then output is set to one else output is set to zero. For every  $p$  there will be  $N$  comparisons and the output for each pixel  $p$  will be of  $N$  bits and that can be presented by decimal number. In practice, the values of  $N$  and  $R$  are 8 and 1, respectively. For a given image or patch, the histogram of *LBP* is computed where the length of histogram is  $2^N$ . Whereas, CSLBP is quantized representation of *LBP*. In *CSLBP*, instead of comparing each neighbor with  $p$ , only center-symmetric neighbors are compared:

$$CSLBP_{N,R,T}(p) = \sum_{i=1}^{\frac{N}{2}} s(|n_i| - |n_{i+\frac{N}{2}}|)2^{i-1}, \quad (1)$$

$$s(j) = \begin{cases} 1 & j > T \\ 0 & \text{otherwise} \end{cases}$$

The length of histogram in CSLBP is  $2^{\frac{N}{2}}$  which is quite shorter than the histogram of *LBP*. The suggested values for  $N, R$  and  $T$  are 8, 1, 0.01, respectively.

To compute the *CSLBP* descriptor, the given patch  $P$  is divided into spatial grid of  $G_x \times G_y$  and the histogram of *CSLBP* is computed for each cell. Finally, all histograms are concatenated to one vector. The length of CSLBP descriptor is  $G_x \times G_y \times 2^{\frac{N}{2}}$ . That is quite often the double of SIFT descriptor. For our experiments the values for  $CSLBP_{N,R,T}$  are  $CSLBP_{8,1,0.01}$ , and the highest efficiency is obtained by keeping  $G_x = 4$  and  $G_y = 4$  that makes *CSLBP* the length of 256.

2) *Scale-Invariant Feature Transform*: The SIFT descriptor is the representation of gradient orientation histograms. To compute the SIFT descriptor, the given patch  $P$  is divided into grid of  $G_x \times G_y$ . In each cell the gradient magnitude,

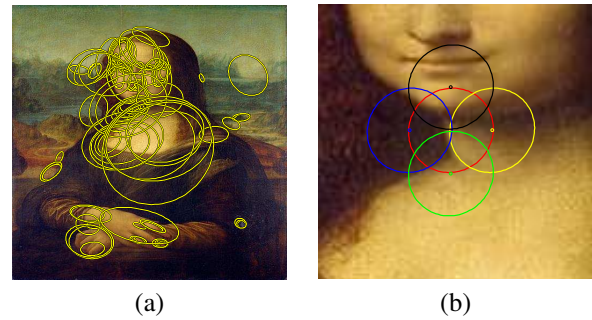


Fig. 2. Neighbor based descriptor computation. (a) shows the keypoints with their elliptical patches, and (b) shows the neighbors of one randomly selected keypoint.

$g(x, y)$ , and orientation,  $\theta(x, y)$ , are computed for each pixel. The gradient orientation are quantized into 8 directions and histogram of quantized orientation is computed. Each sample added to histogram is weighted by its gradient magnitude and Gaussian weight. For Gaussian weight, circular window with a  $\sigma$  that is 1.5 times that of the scale of keypoint is taken [8]. The Gaussian weight is used to give more preference to those pixels that are near to center. Finally, gradient orientation histograms of all cells are concatenated to one vector, SIFT. The maximum efficiency of SIFT is also obtained by keeping the  $G_x = 4$  and  $G_y = 4$ . Therefore, the SIFT descriptor is of 128 length ( $8 \times 4 \times 4$ ).

### III. METHODOLOGY FOR DESCRIPTORS COMPUTATION

In this section, we explain our methodology for descriptors computation which improves the performance for CBCD.

#### A. Feature Extraction and Matching

Local keypoints are extracted and represented as  $q = (x, y, \theta, \sigma, \mathcal{P}_q, \mathbf{d}_q^c)$ , where  $x$  and  $y$  are the coordinates,  $\theta$  is dominant orientation,  $\sigma$  is scale,  $\mathcal{P}$  is 2D affine region of size  $41 \times 41$  centered  $q$ , and  $\mathbf{d}$  is a descriptor vector around keypoint and  $c \in \{CSLBP, SIFT, \dots\}$ .

Two images are said to be similar, or match, if they have many similar descriptors. Two descriptors  $\mathbf{d}_1$  and  $\mathbf{d}_2$  are said to be similar if they are close to each other based on some distance measure. In our framework, we have used Euclidean distance as distance measure. The Euclidean distance is defined as

$$\mathcal{E}(\mathbf{d}_1, \mathbf{d}_2) = \sqrt{\sum_{i=1}^m (\mathbf{d}_1(i) - \mathbf{d}_2(i))^2}, \quad (2)$$

More specifically, given two images  $Q$  and  $R$  with local keypoints sets  $E$  and  $F$ , respectively, we perform nearest neighbor (NN) matching subject to a reliability constraint. We consider the keypoint pair  $(e_i, f_j)$ , where  $e_i \in E, f_j \in F$ , to be similar if their descriptors  $\mathbf{d}_i$  and  $\mathbf{d}_j$  satisfy the following two conditions for distance measure  $\mathcal{E}(\cdot, \cdot)$ :

- Nearest neighbors

$$\mathcal{E}(\mathbf{d}_i, \mathbf{d}_j) = \min_{\mathbf{d}_k \in F} \mathcal{E}(\mathbf{d}_i, \mathbf{d}_k) \quad (3)$$

- Reliable match

$$T_m \cdot \mathcal{E}(\mathbf{d}_i, \mathbf{d}_j) < \min_{\mathbf{d}_l \in F, l \neq j} \mathcal{E}(\mathbf{d}_i, \mathbf{d}_l) \quad (4)$$

where  $T_m > 1$  is a threshold ensuring a stable match under noise conditions.

This method is widely used in computer vision applications [8], [17], [18]. The descriptors with this configuration will be represented by sign D. To make matching fast, we use lookup table in which precomputed distances are stored. In our implementation of descriptors, we used unsigned 8-bits for each element of descriptor vector. Let  $\mathcal{N} = \{0, 1, 2, \dots, 255\}$  be the unsigned 8-bit arrays, then  $\mathcal{T}$  is a function defined as follow:

$$\mathcal{T} : \mathcal{N} \times \mathcal{N} \rightarrow \mathcal{D} \quad (5)$$

where  $\mathcal{D}$  contains the squared difference between two unsigned 8-bits values. Making use of  $\mathcal{T}$  as lookup table, Euclidean distance can be redefined as follow:

$$\mathcal{E}(\mathbf{d}_1, \mathbf{d}_2) = \sqrt{\sum_{i=1}^m \mathcal{T}(\mathbf{d}_1(i), \mathbf{d}_2(i))} \quad (6)$$

Using  $\mathcal{T}$  we can save two mathematical operations, subtraction and squaring (multiplication). We linearly increase the database size of descriptors and find the first nearest neighbor for single descriptor.

### B. Geometrical influenced descriptors computation

For given image, local keypoints are detected. For detected keypoints, affine patches are estimated and normalized and finally respective descriptors are computed, as suggested by Mikolajczyk and Schmid [17]. For keypoints, Harris-affine detector is used which calculates the elliptical like patches for corner like structures [10], [30], [31]. On average, Harris-affine detector returns 1400 points by keeping default parameters on OVG dataset.

For keypoint patch normalization and then descriptor computation, we have used the same steps suggested by Mikolajczyk and Schmid [17].

The performance of descriptors can't be increased by increasing their dimensions with different configuration values of spatial grid ( $G_x$  and  $G_y$ ) nor by increasing the region size around the keypoint (shown in Figure 5). We use different configuration to enrich the descriptor vectors by the gradients or texture in the vicinity. We aim to improve the performance of descriptors before quantization like computation of visual words or codebooks.

1) *Configuration I: Pairing  $K$  spatially close feature descriptors*: In this approach, we use  $K$  nearest neighbor approach. Each keypoint  $q$  is paired with other keypoint descriptor  $r$  which is nearest neighbor of  $q$  in descriptor space. We name this approach nearest neighbor descriptor (NND) paring from now onwards. This approach is sensitive to image noise which are discussed in detail in experimental section.

2) *Configuration II: Increasing the patch size*: To add the spatial information around the keypoint patch, we increase the patch size. The patch size of each keypoint is selected based on their scale to make keypoint scale invariant. Since, the patch  $\mathcal{P}_q$  for given keypoint  $q$  is invariant to scale and affine distortions, therefore, in general scenario the performance of descriptors can be decreased. However, in case of CBCD, the performance is neither increased nor decreased. Experiments show that the performance in case of JPEG compression is increased.

3) *Configuration III: Pairing neighbor region descriptors*: In this configuration, we enrich the keypoint descriptors by the geometrical texture and gradient in vicinity. Instead of increasing the scale or adding an other keypoint descriptor to given keypoint descriptor  $\mathbf{d}_q$ , we take some local image patches near  $\mathcal{P}_q$  and add the gradient or intensity information to  $\mathbf{d}_q$ . This is achieved by taking an other point(s)  $(x', y')$  at the pixel distance  $R_n$  from keypoint  $q$ , the patch over  $(x', y')$  is computed by keeping the same parameters of keypoint  $q$  except  $(x, y)$ , finally newly computed descriptors are concatenated to  $\mathbf{d}_q$ . The example of neighbor local image patches near keypoints is shown in Figure 2. During experiments, we take upto  $\mathcal{N}$  neighbor points, where  $\mathcal{N} = 4$  and name it neighbor based descriptors (NBD). It is not necessarily needed to compute the full length descriptors from neighbor points. We take less number of spatial grid around the patch to compute the descriptors.

The value of  $R_n$  is carefully selected. Since, we are using elliptical regions, so we take *vertex* or *co-vertex* points for the distance of  $R_n$  so that neighbor patches have region overlap with  $\mathcal{P}$  as shown in Figure 2.

## IV. EXPERIMENTS AND RESULTS

To validate the performance of our proposed framework we experiment on two applications. In first one, keypoint matching accuracy provided by homography are computed under different types of image transformations. The transformations mainly include  $C4$ ,  $C5$  and  $C7$ , the list of transformations can be seen in Table I. In second application, the performance is shown for image retrieval application. For image retrieval, query image is matched with all the images and ranked list based on distances is obtained. The images during this experiments have severe transformations which are the mixture of all listed transformations listed in Table I.

### A. Benchmark Datasets

Two different benchmark datasets are used to validate the features. Dataset provided by oxford vision group (OVG)<sup>1</sup> is used for first application experiment. This is standard dataset and used in many papers [17], [18], [32], [33], which comprises of challenging geometric and photometric transformations. Three types of transformations are used in this experiment: illumination change, image blur (also scale change), and JPEG compression. For any given distortion (transformation)

<sup>1</sup><http://www.robots.ox.ac.uk/~vgg/research/affine/>

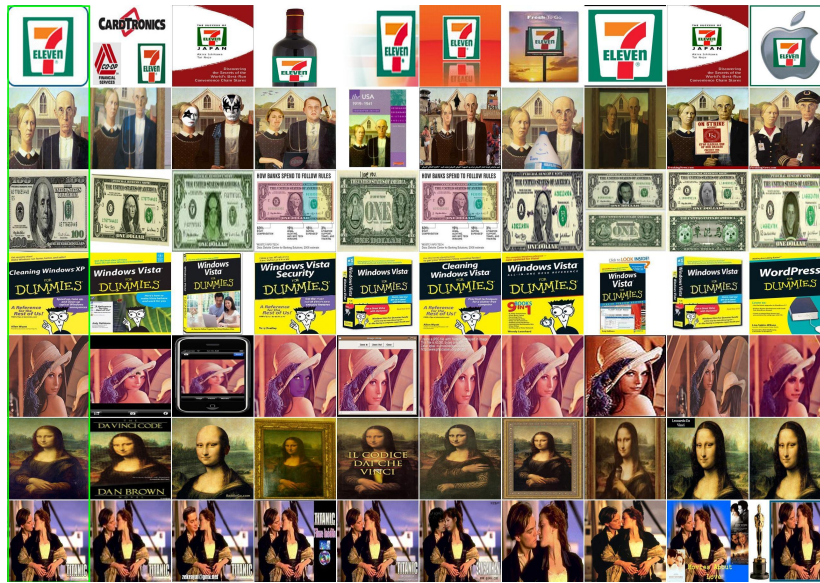


Fig. 3. Dataset used in image-to-image matching for partial image copy detection.

$G$ , base image  $I_0$  of some particular scene is provided with five more gradual distorted images,  $S = \{I_0, I_1, \dots, I_5\}$ . Since, distortion get severe gradually therefore  $I_z$  has less severe transformation compared to  $I_{z+1}$  where  $z \in \{1, 2, \dots, 4\}$ , all images in dataset are related by homography. Detail on acquisition of images and transformations in the dataset can be found in the original work [10], [17].

In second experiment, image retrieval datasets known as PICDD [1], [2] and Oxford [34] are used. In PICDD dataset, the images are collected and manually annotated of different scenes and objects. We randomly select 10 objects/scenes with their 10 transformed copies, the transformations are the union of all transformations listed in Table I. The example of random sample from PICDD dataset is shown in Figure 3. The second dataset, Oxford dataset [34], contains 11 landmarks with 55 query images among total of 5K images obtained from Flickr <sup>2</sup>.

### B. Evaluation matrices

The *precision*, *recall* and *F-score* are used as evaluation matrices

$$\begin{aligned}
 \text{recall} &= \frac{\# \text{ correct matches}}{\text{Total correspondences}} \\
 \text{precision} &= \frac{\# \text{ correct matches}}{\text{Total matches}} \\
 \text{F-score} &= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
 \end{aligned} \tag{7}$$

For point-to-point correspondences, the number of *correct matches* and *total correspondences* are determined by the *overlap error* [10], [17]. The overlap error is the estimation of region overlap in case of transformations, and can easily be estimated with the help of homography, more details can be

<sup>2</sup><http://www.flickr.com/>

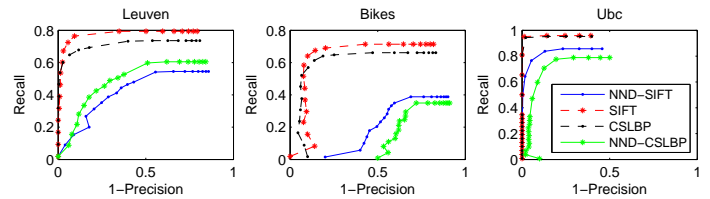


Fig. 4. Performance of NND for Experiment I

found in original work [10], [17]. The values of precision and recall are obtained by the changing the  $T$  which is explained in Section III-A. In case of *F-score*, *precision* is computed on gradually increasing *recall* and only the maximum values of *F-score* are reported. For Experiment II, precision is computed on ranked list. For each image there are 10 copies in database.

### C. Experiment I: Point to point correspondences

In this experiment, local keypoints are detected from images  $\{I_0, I_1, \dots, I_5\}$ . All descriptors SIFT, CSLBP, NBD-SIFT, NBD-CSLBP, NND-SIFT, NND-CSLBP, DSIFT, and DCSLBP are computed from all keypoints in all images. In each transformation, image  $I_0$  is matched with rest of the 5 gradual deformed copies. Due to the page and space limitation, we only report the correct and false matches between one pair images  $I_0$  and  $I_3$ . As explained in previous section that images are related by homography and number correct matches are obtained by overlap error, and due to homography total number of correspondences are already know. For matching we used the methodology explained in Section III-A. We visualize the correct and false matches by recall vs 1-precision, and F-score.

### D. Experiment II: Image retrieval

For Experiment II and dataset PICDD, there are 100 images including 10 query images. For image  $I_Q$ , similarity based matching with all the images in databases are computed. Based



TABLE II  
LIST OF SPATIAL GRID CONFIGURATION USED FOR DESCRIPTOR COMPUTATION. FOR CONF 1 TO CONF 4, NEIGHBORS ARE PROGRESSIVELY INCREASED FROM 1 TO 4.

Type	Description
Conf 1	Spatial grid around keypoint and neighbor is 1 x 2.
Conf 2	Spatial grid around keypoint and neighbor is 2 x 2.
Conf 3	Spatial grid around keypoint is 3 x 3 and grid around neighbors is 2 x 2.
Conf 4	Spatial grid around keypoint and neighbor is 3 x 3.
SIFT / CSLBP	Spatial grid size $\{(2 \times 2), (3 \times 3), \dots, (6 \times 6)\}$
DCSLBP	Double support region around keypoint for CSLBP
DSIFT	Double support region around keypoint for SIFT

on similarity score rank list is maintained. The similarity score between  $I_Q$  and image  $I_d$  is computed as follow

$$\mathcal{M}(I_Q, I_d) = \frac{|I_Q \cup I_d|}{|I_Q|} \times 100 \quad (8)$$

where  $|I_Q \cup I_d|$  represents the similar features between  $I_Q$  and  $I_d$ , and  $|I_Q|$  is total number of features in image  $I_Q$ . The retrieval accuracy is shown in Figure 6. It can be seen that NBD SIFT retrieve, on average, top three images as true positives whereas all other descriptors retrieve only top one as true positive.

We also evaluate the performance of NBD descriptors on larger dataset (Oxford dataset). We follow same steps used in BIGOH [19]. We evaluate the retrieval performance of proposed descriptors by mean average precision (mAP). This is obtained by computing the matching scores of query images with all the images in the database exhaustively. Rank list is obtained for every query image and calculate the precision of retrieval at that cutoff (the number of copies divided by the number of gallery images with match score above the threshold), and finally mean of average precision is calculated. The mAP is shown in Table III.

### E. Results and Discussions

The NND does not add any improvement in descriptors performance as they are very sensitive to image noise. It has been observed that keypoint changes their position in features space over small distortion. The chances to get 1-NN after some image noise is very small. The performance of NND is shown in Figure 4, NND-SIFT and NND-CSLBP are obtained by paring the keypoint descriptors with their first nearest neighbor descriptor. For Configuration II and III, we use different configuration values of spatial grid over patch before descriptors computation. The list of spatial grid configurations is shown in Table II. The different configuration of spatial grid show that descriptors performance can not be improved while increasing the spatial grid size. Whereas, it can be seen that equivalent performance is achieved with

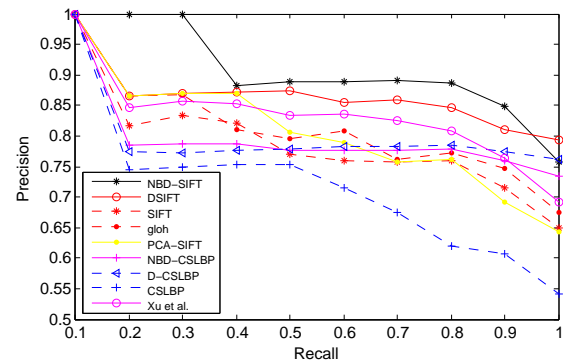


Fig. 6. Performance of SIFT and Neighbor SIFT for image retrieval.

reduced dimensions in the case of NBD descriptors. The descriptors performance for Configuration II and III is shown in Figure 5. For Configuration II, we can see that descriptor performance is not improved. There is only slight different in F-score except JPEG compression, D-CSLBP is better then CSLBP. For Configuration III (neighbor based descriptors), the performance of SIFT and CSLBP is increased. We also get equal performance despite of lower dimensions.

Descriptors computed on many neighbors give better performance, but if the position of keypoint is drastically changed then these descriptors give poor performance. Based on experiments, we recommend to use up to three neighbors. It can be seen that only increasing the scale does not bring any improvement. Instead it improves when descriptor computed from it original patch is concatenated with its neighbors, as affine region around the keypoint have vital contribution for descriptor robustness.

The framework also shows good performance for image retrieval experiment where severely transformed copies of query images are retrieved based on their descriptors matching. Neighbor based descriptors increase the discrimination power and also decrease the ambiguity. The discriminative power of SIFT after enriching it by its neighbor can be seen in Figure 6, where the values of  $G_x$  and  $G_y$  are  $(3 \times 3)$ , and for neighbor patches it is  $2 \times 2$  (only two neighbors are used). The NBD descriptors contains more local information around the keypoint patch therefore when matching score of SIFT and NBD SIFT is computed on true negative images are very different. Ideally, the average matching score with true negative images should be zero but it is not obtained in practical. The average matching score of NBD SIFT is **0.5** whereas for SIFT it is **3.7** on true negative images. This clearly shows that NBD based descriptors have less ambiguity. An example of image-to-image matching based on local keypoints is shown in Figure 1 where it is can be seen that NBD SIFT does not have any matching between different images and have correct matching in case of similar images with symmetric structure.

### V. CONCLUSION

This paper introduces the neighbor based descriptors. Descriptors are influenced by the texture and gradient in vicinity.

TABLE III  
RETRIEVAL ACCURACY (MAP) OF ALL DESCRIPTORS ON LARGER DATASET.

Datasets	SIFT	CSLBP	GLOH	NBD SIFT	NBD CSLBP	D-SIFT	PCA-SIFT
PICDD	0.644	0.510	0.690	0.730	0.590	0.601	0.663
Oxford	0.711	0.602	0.711	0.767	0.610	0.612	0.690

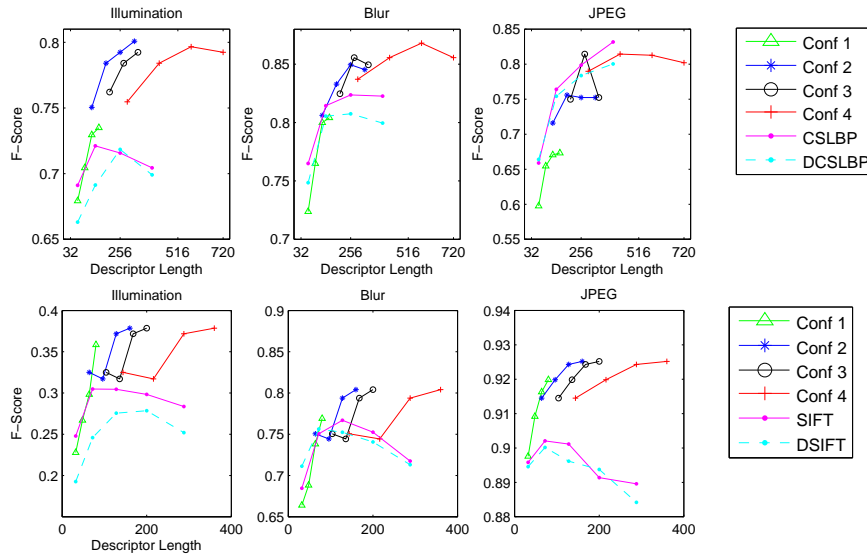


Fig. 5. Results of Keypoint matching for neighbor based descriptors. Top row shows the performance of CSLBP, and bottom row shows of SIFT. Configuration of descriptors computation are shown in Table II. During descriptor computation one to four neighbors are used.

Results on two famous descriptors, SIFT and CSLBP, are shown in Figure 5 and Figure 6. We obtained better performance despite of lower length after influencing the descriptor by their neighbors on particular transformations. Currently, we are trying to explore the retrieval efficiency for very large datasets in real time. Computation of neighbor descriptors increases the computation time. We can also compute hybrid descriptors, adding the power of different descriptors into one. For example, keypoint descriptor can be computed by SIFT and neighbors can be computed by CSLBP or visa-versa.

#### ACKNOWLEDGMENTS

We are thankful to National Institute of Informatics, Tokyo, for supporting the research.

#### REFERENCES

- [1] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Computer Vision and Pattern Recognition*, 2009.
- [2] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, "Spatial coding for large scale partial-duplicate web image search," in *Proceedings of the international conference on Multimedia*, 2010.
- [3] C.-T. Hsu and J.-L. Wu, "Hidden digital watermarks in images," *Image Processing, IEEE Transactions on*, 1999.
- [4] I. Cox, J. Kilian, F. Leighton, and T. Shamon, "Secure spread spectrum watermarking for multimedia," *Image Processing, IEEE Transactions on*, 1997.
- [5] B. Chen and G. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," in *Information Theory, 2000. Proceedings. IEEE International Symposium on*, 2000, p. 46.
- [6] W. Zeng and B. Liu, "A statistical watermark detection technique without using original images for resolving rightful ownerships of digital images," *Image Processing, IEEE Transactions on*, 1999.
- [7] S. Lian, N. Nikolaidis, and H. T. Sencar, "Content-based video copy detection a survey," in *Intelligent Multimedia Analysis for Security Applications*, 2010.
- [8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.
- [9] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, 2008.
- [10] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *IJCV*, 2004.
- [11] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *CVPR*, 2003.
- [12] D. Liu, G. Hua, P. Viola, and T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," in *CVPR*, 2008.
- [13] N. Morioka and S. Satoh, "Building compact local pairwise codebook with joint feature space clustering," in *ECCV*, 2010.
- [14] N. Morioka and S. Satoh, "Compact correlation coding for visual object categorization," in *ICCV*, 2011.
- [15] S. Lazebnik, C. Schmid, and J. Ponce, "A maximum entropy framework for part-based texture and object recognition," in *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01*, 2005.
- [16] J. Baber, S. Satoh, C. Keatmanee, and N. Afzulpurkar, "Improving the performance of sift and cslbp for image copy detection," in *Telecommunications and Signal Processing (TSP), 2013 36th International Conference on*. IEEE, 2013, pp. 803–807.
- [17] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on PAMI*, 2005.
- [18] M. Heikkila, M. Pietikainen, and C. Schmid, "Description of interest regions with local binary patterns," *Pattern Recognition*, 2009.
- [19] J. Baber, M. N. Dailey, S. Satoh, N. Afzulpurkar, and M. Bakhtyar, "BIG-OH: Binarization of gradient orientation histograms," *Image and Vision Computing*, vol. 32, no. 11, pp. 940–953, 2014.
- [20] E. Y. Chang, J. Z. Wang, C. Li, and G. Wiederhold, "RIME: A Replicated Image Detector for the World-Wide Web," in *Storage and Retrieval for Image and Video Databases*, 1998.
- [21] C. Kim, "Content-based image copy detection," *Signal Processing: Image Communication*, vol. 18, no. 3, pp. 169–184, 2003.

- [22] Y. Wan, Q. Yuan, S. Ji, L. He, and Y. Wang, "A survey of the image copy detection," 2008, pp. 738–743.
- [23] A. Basit, M. N. Dailey, J. Moonrinta, and P. Laksanacharoen, "Joint localization and target tracking with a monocular camera," *Robotics and Autonomous Systems*, vol. 74, Part A, pp. 1–14, 2015.
- [24] A. Basit, W. S. Qureshi, M. N. Dailey, and T. Krajník, "Joint localization of pursuit quadcopters and target using monocular cues," *Journal of Intelligent & Robotic Systems*, vol. 78, no. 3, pp. 613–630, 2014. [Online]. Available: <http://dx.doi.org/10.1007/s10846-014-0081-2>
- [25] Z. Xu, H. Ling, F. Zou, Z. Lu, and P. Li, "A novel image copy detection scheme based on the local multi-resolution histogram descriptor," *Multimedia Tools and Applications*, vol. 52, pp. 445–463, 2011.
- [26] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *Proc. CVPR*, 2004, pp. 511–517.
- [27] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," in *ECCV*, 2010.
- [28] M. Heikkilä and M. Pietikäinen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657–662, 2006.
- [29] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*, 11/2011 2011.
- [30] B. Adam, "Reliable feature matching across widely separated views," in *CVPR*, 2000.
- [31] T. Lindeberg and J. Gärding, "Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure," *Image and Vision Computing*, 1997.
- [32] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *ECCV*, 2010.
- [33] M. Ambai and Y. Yoshida, "CARD: Compact And Real-time Descriptors," in *International Conference on Computer Vision*, 2011.
- [34] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.