

Volume 5 Issue 7

July 2014



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)



www.ijacsa.thesai.org



INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS



THE SCIENCE AND INFORMATION ORGANIZATION
www.thesai.org | info@thesai.org

OAlster

getCITED

Google Scholar BETA

BASE
Bielefeld Academic Search Engine

ULRICHSWEB™
GLOBAL SERIALS DIRECTORY

arXiv.org

DOAJ | DIRECTORY OF
OPEN ACCESS
JOURNALS

IET InspecDirect

INDEX COPERNICUS
INTERNATIONAL

WorldCat
Window to the world's libraries

Microsoft Academic Search
Beta

EBSCO
HOST
Research
Databases

Editorial Preface

From the Desk of Managing Editor...

It is our pleasure to present to you the July 2014 Issue of International Journal of Advanced Computer Science and Applications.

Today, it is incredible to consider that in 1969 men landed on the moon using a computer with a 32-kilobyte memory that was only programmable by the use of punch cards. In 1973, Astronaut Alan Shepherd participated in the first computer "hack" while orbiting the moon in his landing vehicle, as two programmers back on Earth attempted to "hack" into the duplicate computer, to find a way for Shepherd to convince his computer that a catastrophe requiring a mission abort was not happening; the successful hack took 45 minutes to accomplish, and Shepherd went on to hit his golf ball on the moon. Today, the average computer sitting on the desk of a suburban home office has more computing power than the entire U.S. space program that put humans on another world!!

Computer science has affected the human condition in many radical ways. Throughout its history, its developers have striven to make calculation and computation easier, as well as to offer new means by which the other sciences can be advanced. Modern massively-paralleled super-computers help scientists with previously unfeasible problems such as fluid dynamics, complex function convergence, finite element analysis and real-time weather dynamics.

At IJACSA we believe in spreading the subject knowledge with effectiveness in all classes of audience. Nevertheless, the promise of increased engagement requires that we consider how this might be accomplished, delivering up-to-date and authoritative coverage of advanced computer science and applications.

Throughout our archives, new ideas and technologies have been welcomed, carefully critiqued, and discarded or accepted by qualified reviewers and associate editors. Our efforts to improve the quality of the articles published and expand their reach to the interested audience will continue, and these efforts will require critical minds and careful consideration to assess the quality, relevance, and readability of individual articles.

To summarise, the journal has offered its readership thought provoking theoretical, philosophical, and empirical ideas from some of the finest minds worldwide. We thank all our readers for their continued support and goodwill for IJACSA. We will keep you posted on updates about the new programmes launched in collaboration.

Lastly, 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.

We hope that materials contained in this volume will satisfy your expectations and entice you to submit your own contributions in upcoming issues of IJACSA

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 5 Issue 7 July 2014
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: Human-Computer Interaction, Networking, Information Retrievals, Optimization Theory, Modelling and Simulation, Satellite Remote Sensing, Computer Vision, Decision Making Methodology

Associate Editors

Chao-Tung Yang

Department of Computer Science, Tunghai University, Taiwan

Domain of Research: Cloud Computing

Elena SCUTELNICU

"Dunarea de Jos" University of Galati, Romania

Domain of Research: e-Learning Tools, Modelling and Simulation of Welding Processes

Krassen Stefanov

Professor at Sofia University St. Kliment Ohridski, Bulgaria

Domains of Research: Digital Libraries

Maria-Angeles Grado-Caffaro

Scientific Consultant, Italy

Domain of Research: Sensing and Sensor Networks

Mohd Helmy Abd Wahab

Universiti Tun Hussein Onn Malaysia

Domain of Research: Data Mining, Database, Web-based Application, Mobile Computing

T. V. Prasad

Lingaya's University, India

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

Reviewer Board Members

- **Abbas Karimi**
Islamic Azad University Arak Branch
- **Abdel-Hameed Badawy**
Arkansas Tech University
- **Abdelghni Lakehal**
Fsdm Sidi Mohammed Ben Abdellah University
- **Abeer Elkorny**
Faculty of computers and information, Cairo University
- **ADEMOLA ADESINA**
University of the Western Cape, South Africa
- **Ahmed Boutejdar**
- **Dr. Ahmed Nabih Zaki Rashed**
Menoufia University
- **Aderemi A. Atayero**
Covenant University
- **Akbar Hossin**
- **Akram Belghith**
University Of California, San Diego
- **Albert S**
Kongu Engineering College
- **Alcinia Zita Sampaio**
Technical University of Lisbon
- **Ali Ismail Awad**
Luleå University of Technology
- **Alexandre Bouënard**
- **Amitava Biswas**
Cisco Systems
- **Anand Nayyar**
KCL Institute of Management and Technology, Jalandhar
- **Andi Wahju Rahardjo Emanuel**
Maranatha Christian University, INDONESIA
- **Anirban Sarkar**
National Institute of Technology, Durgapur, India
- **Anuranjan misra**
Bhagwant Institute of Technology, Ghaziabad, India
- **Andrews Samraj**
Mahendra Engineering College
- **Arash Habibi Lashakri**
University Technology Malaysia (UTM)
- **Aris Skander**
Constantine University
- **Ashraf Mohammed Iqbal**
Dalhousie University and Capital Health
- **Ashok Matani**
- **Ashraf Owis**
Cairo University
- **Asoke Nath**
St. Xaviers College
- **Ayad Ismaeel**
Department of Information Systems Engineering- Technical Engineering College-Erbil / Hawler Polytechnic University, Erbil-Kurdistan Region- IRAQ
- **Babatunde Opeoluwa Akinkunmi**
University of Ibadan
- **Badre Bossoufi**
University of Liege
- **Basil Hamed**
Islamic University of Gaza
- **Bharti Waman Gawali**
Department of Computer Science & information
- **Bhanu Prasad Pinnamaneni**
Rajalakshmi Engineering College; Matrix Vision GmbH
- **Bilian Song**
LinkedIn
- **Brahim Raouyane**
FSAC
- **Brij Gupta**
University of New Brunswick
- **Bright Keswani**
Suresh Gyan Vihar University, Jaipur (Rajasthan) INDIA
- **Constantin Filote**
Stefan cel Mare University of Suceava
- **Constantin Popescu**
Department of Mathematics and Computer Science, University of Oradea
- **Chandrashekhar Meshram**
Chhattisgarh Swami Vivekananda Technical University
- **Chao Wang**

- **Chi-Hua Chen**
National Chiao-Tung University
- **Ciprian Dobre**
University Politehnica of Bucharest
- **Chien-Pheg Ho**
Information and Communications Research
Laboratories, Industrial Technology Research
Institute of Taiwan
- **Charlie Obimbo**
University of Guelph
- **Chao-Tung Yang**
Department of Computer Science, Tunghai
University
- **Dana PETCU**
West University of Timisoara
- **Deepak Garg**
Thapar University
- **Dewi Nasien**
Universiti Teknologi Malaysia
- **Dheyaa Kadhim**
University of Baghdad
- **Dong-Han Ham**
Chonnam National University
- **Dragana Becejski-Vujaklija**
University of Belgrade, Faculty of organizational
sciences
- **Driss EL OUADGHIRI**
- **Duck Hee Lee**
Medical Engineering R&D Center/Asan Institute for
Life Sciences/Asan Medical Center
- **Dr. Santosh Kumar**
Graphic Era University, Dehradun, India
- **Elena Camossi**
Joint Research Centre
- **Eui Lee**
- **Elena SCUTELNICU**
"Dunarea de Jos" University of Galati
- **Firkhan Ali Hamid Ali**
UTHM
- **Fokrul Alom Mazarbhuiya**
King Khalid University
- **Frank Ibikunle**
Covenant University
- **Fu-Chien Kao**
Da-Y eh University
- **Faris Al-Salem**
- GCET
- **gamil Abdel Azim**
Associate prof - Suez Canal University
- **Ganesh Sahoo**
RMRIMS
- **Gaurav Kumar**
Manav Bharti University, Solan Himachal Pradesh
- **Ghalem Belalem**
University of Oran (Es Senia)
- **Giri Babu**
Indian Space Research Organisation
- **Giacomo Veneri**
University of Siena
- **Giri Babu**
Indian Space Research Organisation
- **Gerard Dumancas**
Oklahoma Medical Research Foundation
- **Georgios Galatas**
- **George Mastorakis**
Technological Educational Institute of Crete
- **Gunaseelan Devaraj**
Jazan University, Kingdom of Saudi Arabia
- **Gavril Grebenisan**
University of Oradea
- **Hadj Tadjine**
IAV GmbH
- **Hamid Mukhtar**
National University of Sciences and Technology
- **Hamid Alinejad-Rokny**
University of Newcastle
- **Harco Leslie Hendric Spits Warnars**
Budi LUhur University
- **Harish Garg**
Thapar University Patiala
- **Hamez I. El Shekh Ahmed**
Pure mathematics
- **Hesham Ibrahim**
Chemical Engineering Department, Faculty of
Engineering, Al-Mergheb University
- **Dr. Himanshu Aggarwal**
Punjabi University, India
- **Huda K. AL-Jobori**
Ahlia University
- **Iwan Setyawan**
Satya Wacana Christian University

- **Dr. Jamaiah Haji Yahaya**
Northern University of Malaysia (UUM), Malaysia
- **James Coleman**
Edge Hill University
- **Jim Wang**
The State University of New York at Buffalo,
Buffalo, NY
- **John Salin**
George Washington University
- **Jyoti Chaudary**
High performance computing research lab
- **Jatinderkumar R. Saini**
S.P.College of Engineering, Gujarat
- **K Ramani**
K.S.Rangasamy College of Technology,
Tiruchengode
- **K V.L.N.Acharyulu**
Bapatla Engineering college
- **Kashif Nisar**
Universiti Utara Malaysia
- **Kayhan Zrar Ghafoor**
University Technology Malaysia
- **Kitimaporn Choochote**
Prince of Songkla University, Phuket Campus
- **Kunal Patel**
Ingenuity Systems, USA
- **Krasimir Yordzhev**
South-West University, Faculty of Mathematics and
Natural Sciences, Blagoevgrad, Bulgaria
- **Krassen Stefanov**
Professor at Sofia University St. Kliment Ohridski
- **Labib Francis Gergis**
Misr Academy for Engineering and Technology
- **Lai Khin Wee**
Biomedical Engineering Department, University
Malaya
- **Lazar Stosic**
Collegefor professional studies educators Aleksinac,
Serbia
- **Lijian Sun**
Chinese Academy of Surveying and Mapping, China
- **Leandors Maglaras**
- **Leon Abdillah**
Bina Darma University
- **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**
- **M. Tariq Banday**
University of Kashmir
- **MAMTA BAHETI**
SNJBS KBJ COLLEGE OF ENGINEERING, CHANDWAD,
NASHIK, M.S. INDIA
- **Mazin Al-Hakeem**
Research and Development Directorate - Iraqi
Ministry of Higher Education and Research
- **Md Rana**
University of Sydney
- **Miriampally Venkata Raghavendera**
Adama Science & Technology University, Ethiopia
- **Mirjana Popvic**
School of Electrical Engineering, Belgrade University
- **Manas deep**
Masters in Cyber Law & Information Security
- **Manpreet Singh Manna**
SLIET University, Govt. of India
- **Manuj Darbari**
BBD University
- **Md. Zia Ur Rahman**
Narasaraopeta Engg. College, Narasaraopeta
- **Messaouda AZZOUZI**
Ziane AChour University of Djelfa
- **Dr. Michael Watts**
University of Adelaide
- **Milena Bogdanovic**
University of Nis, Teacher Training Faculty in Vranje
- **Miroslav Baca**
University of Zagreb, Faculty of organization and
informatics / Center for biomet
- **Mohamed Ali Mahjoub**
Preparatory Institute of Engineer of Monastir
- **Mohamed El-Sayed**
Faculty of Science, Fayoum University, Egypt
- **Mohammad Yamin**
- **Mohammad Ali Badamchizadeh**
University of Tabriz
- **Mohamed Najeh Lakhoua**
ESTI, University of Carthage

- **Mohammad Alomari**
Applied Science University
- **Mohammad Kaiser**
Institute of Information Technology
- **Mohammed Al-Shabi**
Assistant Prof.
- **Mohammed Sadgal**
- **Mourad Amad**
Laboratory LAMOS, Bejaia University
- **Mohammed Ali Hussain**
Sri Sai Madhavi Institute of Science & Technology
- **Mohd Helmy Abd Wahab**
Universiti Tun Hussein Onn Malaysia
- **Mueen Uddin**
Universiti Teknologi Malaysia UTM
- **Mona Elshinawy**
Howard University
- **Maria-Angeles Grado-Caffaro**
Scientific Consultant
- **Mehdi Bahrami**
University of California, Merced
- **Miriampally Venkata Raghavendra**
Adama Science & Technology University, Ethiopia
- **Murthy Dasika**
SreeNidhi Institute of Science and Technology
- **Mostafa Ezziyani**
FSTT
- **Marcellin Julius Nkenlifack**
University of Dschang
- **Natarajan Subramanyam**
PES Institute of Technology
- **Noura Aknin**
University Abdelamlek Essaadi
- **Nidhi Arora**
M.C.A. Institute, Ganpat University
- **Nazeeruddin Mohammad**
Prince Mohammad Bin Fahd University
- **Najib Kofahi**
Yarmouk University
- **NEERAJ SHUKLA**
ITM UNiversity, Gurgaon, (Haryana) Inida
- **N.Ch. Iyengar**
VIT University
- **Om Sangwan**
- **Oliviu Matel**
Technical University of Cluj-Napoca
- **Osama Omer**
Aswan University
- **Ousmane Thiare**
Associate Professor University Gaston Berger of
Saint-Louis SENEGAL
- **Omaima Al-Allaf**
Assistant Professor
- **Paresh V Virparia**
Sardar Patel University
- **Dr. Poonam Garg**
Institute of Management Technology, Ghaziabad
- **Professor Ajantha Herath**
- **Prabhat K Mahanti**
UNIVERSITY OF NEW BRUNSWICK
- **Qufeng Qiao**
University of Virginia
- **Rachid Saadane**
EE departement EHTP
- **raed Kanaan**
Amman Arab University
- **Raja boddu**
LENORA COLLEGE OF ENGINEERNG
- **Ravisankar Hari**
SENIOR SCIENTIST, CTRI, RAJAHMUNDRY
- **Raghuraj Singh**
- **Rajesh Kumar**
National University of Singapore
- **Rakesh Balabantaray**
IIIT Bhubaneswar
- **RashadAl-Jawfi**
Ibb university
- **Rashid Sheikh**
Shri Venkateshwar Institute of Technology , Indore
- **Ravi Prakash**
University of Mumbai
- **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, Tasmaniia, Australia
- **Rowayda Sadek**
- **Ruchika Malhotra**
Delhi Technoogical University
- **Saadi Slami**
University of Djelfa

- **Sachin Kumar Agrawal**
University of Limerick
- **Dr.Sagarmay Deb**
University Lecturer, Central Queensland University,
Australia
- **Said Ghoniemy**
Taif University
- **Sasan Adibi**
Research In Motion (RIM)
- **Sérgio Ferreira**
School of Education and Psychology, Portuguese
Catholic University
- **Sebastian Marius Rosu**
Special Telecommunications Service
- **Selem charfi**
University of Valenciennes and Hainaut Cambresis,
France.
- **Seema Shah**
Vidyalankar Institute of Technology Mumbai,
- **Sengottuvelan P**
Anna University, Chennai
- **Senol Piskin**
Istanbul Technical University, Informatics Institute
- **Seyed Hamidreza Mohades Kasaei**
University of Isfahan
- **Shafiqul Abidin**
G GS I P University
- **Shahanawaj Ahamad**
The University of Al-Kharj
- **Shawkl Al-Dubae**
Assistant Professor
- **Shriram Vasudevan**
Amrita University
- **Sherif Hussain**
Mansoura University
- **Siddhartha Jonnalagadda**
Mayo Clinic
- **Sivakumar Poruran**
SKP ENGINEERING COLLEGE
- **Sim-Hui Tee**
Multimedia University
- **Simon Ewedafe**
Baze University
- **SUKUMAR SENTHILKUMAR**
Universiti Sains Malaysia
- **Slim Ben Saoud**
- **Sudarson Jena**
GITAM University, Hyderabad
- **Sumit Goyal**
- **Sumazly Sulaiman**
Institute of Space Science (ANGKASA), Universiti
Kebangsaan Malaysia
- **Sohail Jabb**
Bahria University
- **Suhas J Manangi**
Microsoft
- **Suresh Sankaranarayanan**
Institut Teknologi Brunei
- **Susarla Sastry**
J.N.T.U., Kakinada
- **Syed Ali**
SMI University Karachi Pakistan
- **T C. Manjunath**
HKBK College of Engg
- **T V Narayana Rao**
Hyderabad Institute of Technology and
Management
- **T. V. Prasad**
Lingaya's University
- **Taiwo Ayodele**
Infonetmedia/University of Portsmouth
- **Tarek Gharib**
- **THABET SLIMANI**
College of Computer Science and Information
Technology
- **Totok R. Biyanto**
Engineering Physics, ITS Surabaya
- **TOUATI YOUCEF**
Computer sce Lab LIASD - University of Paris 8
- **VINAYAK BAIRAGI**
Sinhgad Academy of engineering, Pune
- **VISHNU MISHRA**
SVNIT, Surat
- **Vitus S.W. Lam**
The University of Hong Kong
- **Vuda SREENIVASARAO**
School of Computing and Electrical
Engineering,BAHIR DAR UNIVERSITY, BAHIR
DAR,ETHIOPA
- **Vaka MOHAN**
TRR COLLEGE OF ENGINEERING
- **Wei Wei**
- **Xiaoqing Xiang**
AT&T Labs

- **YASSER ATTIA ALBAGORY**
College of Computers and Information Technology,
Taif University, Saudi Arabia
- **YI FEI WANG**
The University of British Columbia
- **Yilun Shang**
University of Texas at San Antonio
- **YU QI**
Mesh Capital LLC
- **Zacchaeus Omogbadegun**
Covenant University
- **ZAIRI ISMAEL RIZMAN**

- UiTM (Terengganu) Dungun Campus
- **ZENZO POLITE NCUBE**
North West University
 - **ZHAO ZHANG**
Deptment of EE, City University of Hong Kong
 - **ZHIXIN CHEN**
ILX Lightwave Corporation
 - **ZLATKO STAPIC**
University of Zagreb
 - **Ziyue Xu**
 - **ZURAINI ISMAIL**
Universiti Teknologi Malaysia

CONTENTS

Paper 1: Benefits Management of Cloud Computing Investments

Authors: Richard Greenwell, Xiaodong Liu, Kevin Chalmers

PAGE 1 – 9

Paper 2: Image Segmentation Via Color Clustering

Authors: Kaveh Heidary

PAGE 10 – 16

Paper 3: A Framework to Improve Communication and Reliability Between Cloud Consumer and Provider in the Cloud

Authors: Vivek Sridhar

PAGE 17 – 21

Paper 4: A Wavelet-Based Approach for Ultrasound Image Restoration

Authors: Mohammed Tarek GadAllah, Samir Mohammed Badawy

PAGE 22 – 29

Paper 5: A Second Correlation Method for Multivariate Exchange Rates Forecasting

Authors: Agus Sihabuddin, Subanar, Dedi Rosadi, Edi Winarko

PAGE 30 – 33

Paper 6: Mitigation of Cascading Failures with Link Weight Control

Authors: Hoang Anh Tran Quang, Akira Namatame

PAGE 34 – 40

Paper 7: Applicability of the Maturity Model for IT Service Outsourcing in Higher Education Institutions

Authors: Victoriano Valencia García, Dr. Eugenio J. Fernández Vicente, Dr. Luis Usero Aragonés

PAGE 41 – 50

Paper 8: Modification of CFCM in The Presence of Heavy AWGN for Bayesian Blind Channel Equalizer

Authors: Changkyu Kim, Soowhan Han

PAGE 51 – 58

Paper 9: An Object-Oriented Smartphone Application for Structural Finite Element Analysis

Authors: B.J. Mac Donald

PAGE 59 – 66

Paper 10: New Approach for Image Fusion Based on Curvelet Approach

Authors: Gehad Mohamed Taher, Mohamed ElSayed Wahed, Ghada EL Taweal

PAGE 67 – 73

Paper 11: Automated Menu Recommendation System Based on Past Preferences

Authors: Daniel Simon Sanz, Ankur Agrawal

PAGE 74 – 77

Paper 12: A Shape Based Image Search Technique

Authors: Aratrika Sarkar, PallabiBhatttcharjee

PAGE 78 – 82

Paper 13: Computer Ethics in the Semantic Web Age

Authors: Aziz Alotaibi

PAGE 83 – 85

Paper 14: A Tool Design of Cobit Roadmap Implementation

Authors: Karim Youssfi, Jaouad Boutahar, Souhail Elghazi

PAGE 86 – 94

Paper 15: Ontology Mapping of Business Process Modeling Based on Formal Temporal Logic

Authors: Irfan Chishti, Jixin Ma, Brian Knight

PAGE 95 – 104

Paper 16: Adaptive Cache Replacement:A Novel Approach

Authors: Sherif Elfayoumy, Sean Warden

PAGE 105 – 111

Paper 17: Application of Fuzzy Self-Optimizing Control Based on Differential Evolution Algorithm for the Ratio of Wind to Coal Adjustment of Boiler in the Thermal Power Plant

Authors: Ting Hou, Liping Zhang, Yuchen Chen

PAGE 112 – 116

Paper 18: Automatic Optic Disc Boundary Extraction from Color Fundus Images

Authors: Thresiamma Devasia, Paulose Jacob, Tessamma Thomas

PAGE 117 – 124

Paper 19: Feature Descriptor Based on Normalized Corners and Moment Invariant for Panoramic Scene Generation

Authors: Kawther Abbas Sallal, Abdul-Monem Saleh Rahma

PAGE 125 – 131

Paper 20: Hybrid Client Side Phishing Websites Detection Approach

Authors: Firdous Kausar, Bushra Al-Otaibi, Asma Al-Qadi, Nwayer Al-Dossari

PAGE 132 – 140

Paper 21: A Study of Scala Repositories on Github

Authors: Ron Coleman, Matthew A. Johnson

PAGE 141 – 148

Paper 22: A Crypto-Steganography: A Survey

Authors: Md. Khalid Imam Rahmani, Kamiya Arora, Naina Pal

PAGE 149 – 155

Paper 23: An Ecn Approach to Congestion Control Mechanisms in Mobile Adhoc Networks

Authors: Som Kant Tiwari, Dr.Y.K.Rana, Prof. Anurag Jain

PAGE 156 – 159

Paper 24: Clustering of Image Data Using K-Means and Fuzzy K-Means

Authors: Md. Khalid Imam Rahmani, Naina Pal, Kamiya Arora

PAGE 160 – 163

Paper 25: Identifying and Extracting Named Entities from Wikipedia Database Using Entity Infoboxes

Authors: Muhidin Mohamed, Mourad Oussalah

PAGE 164 – 169

Paper 26: Design and Implementation of an Interpreter Using Software Engineering Concepts

Authors: Fan Wu, Hira Narang, Miguel Cabral

PAGE 170 – 177

Paper 27: A parallel line sieve for the GNFS Algorithm

Authors: Sameh Daoud, Ibrahim Gad

PAGE 178 – 185

Paper 28: Estimating the Number of Test Workers Necessary for a Software Testing Process Using Artificial Neural Networks

Authors: Alaa F. Shefa, Sofian Kassaymeh, David Rine

PAGE 186 – 192

Paper 29: Natural Gradient Descent for Training Stochastic Complex-Valued Neural Networks

Authors: Tohru Nitta

PAGE 193 – 198

Benefits Management of Cloud Computing Investments

Richard Greenwell, Xiaodong Liu and Kevin Chalmers
Institute for Informatics and Digital Innovation
Edinburgh Napier University
UK

Abstract—This paper examines investments in cloud computing using the Benefits Management approach. The major contribution of the paper is to provide a unique insight into how organizations derive value from cloud computing investments. The motivation for writing this paper is to consider the business benefits generated from utilizing cloud computing in a range of organizations. Case studies are used to describe a number of organizations approaches to benefits exploitation using cloud computing. It was found that smaller organizations can generate rapid growth using strategies based on cloud computing. Larger organizations have used utility approaches to reduce the costs of IT infrastructure.

Keywords—Cloud Computing; Benefits Management; Information Systems Management

I. INTRODUCTION

Cloud computing is becoming a key component in many organizations information systems strategy. The motivation for this paper is to provide a framework for organizations to manage value from cloud investments and, to consider a Benefits Management approach towards cloud computing investments. Ward and Daniel[1] have identified high levels of dissatisfaction with the benefits derived from IT/IS projects, as shown in the table below.

TABLE I. DISSATISFACTION LEVELS WITH BENEFITS DERIVED FROM IS/IT ACTIVITIES

| Benefits Management activity | Level of dissatisfaction |
|--|--------------------------|
| Identification of project costs | 43% |
| Project prioritization | 59% |
| Identify benefits | 68% |
| Development of business cases | 69% |
| Planning the delivery of benefits | 75% |
| Evaluation and review of benefits realized | 81% |

The major contribution from the paper is a unique examination of cloud computing as IS/IT investments in an analytical framework of the Benefits Management approach.

A number of case studies provide an insight into how organization of different sizes and types of use cloud

computing. The investigation leads on to future work to improve the Benefits Management approach.

Cloud computing is similar to the time sharing computer services that were prevalent in IS systems in the 1960's and 1970's [2]. Grossman [3] defines cloud computing as “*clouds, or clusters of distributed computers, providing on-demand resources and services over a network, usually the Internet, with the scale and reliability of a data center*”.

Organizations can use combinations of hardware and software as required to deliver IS/IT services with some outsourced provision if required.

A number of provision models for cloud computing exist, which are developed from the NIST standards [4]:

- Infrastructure as a Service (IaaS) - Fundamental computing resources such as hardware and software managed by another party
- Platform as a Service (PaaS) - The capability to deploy applications (created or acquired), which operate on infrastructure managed by another party
- Software as a Service (SaaS) - The ability to use applications from a number of devices managed or controlled by another party.

The models have common features such as elasticity of usage, flexibility of information storage and user self-service. The differences in the models also influence the view of cloud computing, for example IaaS and some SaaS can be seen as utilities, purchased on price PaaS and some SaaS can be used for business transformation.

A number of cloud ownership models have been observed, again defined by the NIST standards [4]:

- Public clouds are cloud infrastructure provided for open use by organization on their premises
- Private clouds are cloud infrastructure that is provisioned for exclusive use by a single organization comprising multiple consumers; that are owned, managed and operated by an organization or other parties.
- Hybrid clouds use a combination of public and private infrastructures, bound by some technology, that enables data and application portability

Cloud computing can be linked to a number of strategic innovations, such as Big Data [5] and Data Science [6]. Low cost ubiquitous cloud computing resources can be used to process information from large datasets or databases (Big Data) and complex statistical and machine learning techniques can be applied to the data sets (Data Science).

This paper examines how a number of organizations deliver value from cloud computing investments. The methodology for the case study selection was to deliberately select organizations of different sizes and types to obtain the maximum number of IS/IT enablers and benefit types.

The paper will continue with an examination of related work on obtaining value from cloud computing Investments. A description of the Benefits Management approach pioneered by Ward and Daniel and others will follow. Four detailed case studies will then be described, followed by analysis and discussion of the case studies. Conclusions and future work complete the paper.

II. RELATED WORK

Ward and Daniel [1] have identified dissatisfaction with current approaches to obtaining value from IT/IS investments. The reasons they give are a focus on technology delivery and concentration on monetary measures. A number of researchers have examined value derived from cloud computing, a discussion now follows.

Moreno-Vozmediano et al. [7] focus on the technology delivery of cloud benefits. They emphasize the operational aspects of cloud management such as scalability, elasticity, security and aggregation of cloud services. Technological enablers (such as cloud investments) are not related to possible change in the business that could deliver benefits.

Low et al [8] take a multi-factor approach to cloud computing adoption and describe relative advantage, top management support, firm size and internal and external pressure as being determining factors in adoption. The paper links relative advantage in technology directly influencing cloud adoption without any connection to change or business benefits.

Misra and Mondal [9] have built a wide ranging model for cloud adoption, based on Return on Investment (ROI). The model is comprehensive and based on a number of variables such cloud resource availability, usage patterns and criticality of work the organisation carries out. This approach does not examine how benefits are delivered to businesses outside the narrow calculation of ROI. Han [10] follows in a similar vein with calculations of Total Cost of Ownership (TCO) based on fixed and variable infrastructure costs. It is important to examine costs and return on investment, as these are key factors in the adoption of cloud computing, however, stakeholders have a wider range of enablers, changes and benefits that must be examined.

Mohammed et al. [11] describe a cloud value chain reference model, based on the well-known value chain technique pioneered by Porter [12]. Monetary values are attached to cloud and business services that use the cloud services. The method concentrates solely on the monetary

aspects of cloud computing adoption and fails to show linkages between technology enablers, change and benefits.

The typical examples described above concentrate on technology delivery or monetary measures of value from cloud computing investments. There is little linkage between technology enablers, change and benefits. A discussion the Benefits Management approach pioneered by researchers such as Ward and Daniel [1] now follows.

III. THE BENEFITS MANAGEMENT APPROACH

Cloud computing can be viewed as an enabling IS/IT innovation and, cannot be seen as creating value or benefit in isolation. Organizations must derive benefit from the innovation and employ a mechanism to create change and business benefit within the organization. This creates a portfolio of investments within the organization that generate competencies and long term competitive advantage. The research question proposed in this paper centers around deriving benefits from technology enablers. An effective Benefits Management system must be in place to generate these benefits.

The Benefits Management approach was developed by Peppard, Daniel, Ward and Rylander [13]; Peppard and Rylander [14]; the researchers developed the approach from empirical studies of IT projects. They found that IT investments failed to deliver benefits to organizations and they aimed to address this shortfall by concentrating on maximising benefits from IT investments. The researchers argue that traditional project and investment approaches force IS project managers to overstate benefits to allow investment to take place [15]. Real benefits are ill-defined or overstated.

To overcome the issues with traditional approaches, benefits must be accurately identified and a plan must be in place to realise the benefits. Ward and Daniel [15] define the Benefits Management approach as “*The process of organising and managing so that the potential benefits from IT are actually realised*”. Benefits Management brings together a number of techniques such as benefits realisation and change management. Peppard et al. [13] describe the benefits realisation approach as viewing IT as providing no inherent value and, with benefits only arising “*When IT enables people do things differently*”. Stakeholders such as business managers and users can realise business benefits which must be actively managed.

Technological enablers drive the benefits process, however, they are converted into benefits by a number of individuals, groups, processes and techniques that trigger change.

Stakeholders (users and business managers) are key to benefits realization. [13]. A criticism of this view is that it ignores large groups of stakeholders within organizations [16]. Start-up companies can be technology led and, ignoring technology stakeholders could lead to potential benefits being lost. The agile approach to software development is carried out by multi-disciplinary teams, many of whom are technologists [17]. The Benefits Management process must capture all benefits and exploit them to be successful. The core tool for

Benefits Management is the Benefits Dependency Network (BDN) [18] which is shown below.

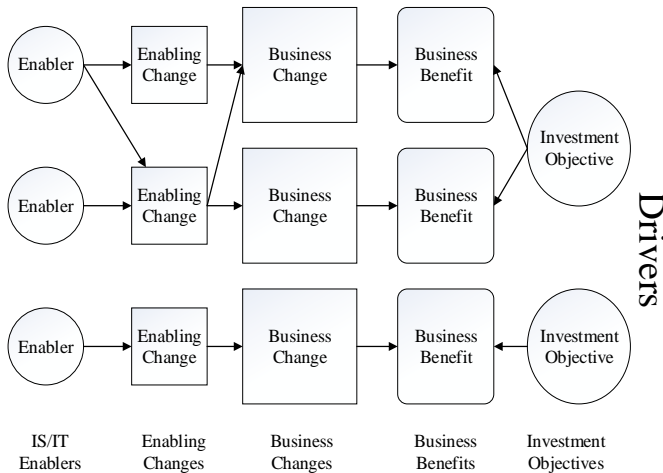


Fig. 1. Benefits Dependency Network

The BDN allows investment objectives and linked benefits to be structured, so that benefits can be realized. The BDN sees enablers generating enabling change that in turn drives business change which creates business benefits. Business benefits are linked to investment objectives in the organization. Enablers are IS/IT technology that cannot create change in isolation. Enabling changes are created through activities such as training or education which trigger business changes that generate benefits that meet investment objectives.

The table below shows the classification of benefits from the BDN. The benefits are described in terms of their degree of explicitness with financial being the most explicit, followed by quantifiable, measurable and observable benefits. An action describes future conduct for the benefit. Doing new things or doing things better with benefits and stop doing things for dis-benefits. The structuring of benefits builds on earlier work on situational analysis, for example SWOT and TOWS [19] and five forces analysis [12].

TABLE II. CLASSIFICATION OF BENEFITS MATRIX

| Degree of Explicitness/Action | Do New Things | Do Things Better | Stop Doing Things |
|-------------------------------|---------------|------------------|-------------------|
| Financial | | | |
| Quantifiable | | | |
| Measurable | | | |
| Observable | | | |

Developing a portfolio of investments allows benefits and risks to be balanced and managed effectively. An example of some aspects of cloud computing previously identified are shown in the table below. The headings of the table show groupings of high potential, strategic, key operational investments and support investments, which will now be described.

TABLE III. PORTFOLIO OF INVESTMENTS FOR CLOUD COMPUTING

| Strategic Investments | High Potential Investments |
|----------------------------------|---|
| SaaS Hybrid Cloud Big Data | PaaS Private Cloud Data Science Human Intelligence Markets |
| Public Cloud | IaaS |
| Key Operation Investments | Support Investments |

High potential investments provide high levels of benefits but also have high risks, and thus must be controlled carefully, except in small start-up organizations that are based around such investments. Organizations can develop new products on cloud based PaaS. There is a development risk as new software has to be created and a large financial commitment in staff and capability development is required.

Private clouds can provide a unique competency for the organization. Google and Amazon [20] have developed their own private clouds to drive their businesses. Data Science is an emerging technology that can provide significant competitive advantage, which can use cloud computing.

Strategic investments are those which require significant change over a medium to long term. Public authorities are considering SaaS solutions to reduce the cost of ownership and to outsource non-key competencies under schemes such as G-Cloud [21]. Hybrid cloud solutions which combine both public and private clouds are key strategic investments. Big Data using cloud based technology is seen as more established and pervasive than technologies such as Data Science [22].

Key operational investments improve systems and services that are critical to the organization. These investments must be tightly controlled and be low risk. Public clouds provide off the shelf computing resources that have been tested by providers. A high degree of redundancy and availability are built into public cloud offerings. Many large organizations are moving key operational systems [23] to public clouds, to provide costs savings and improve availability and reliability of systems. There are risks in loss of control and the security of intellectual property.

Support investments are low risk investments in essential systems. These investments concentrate on efficiencies and cost reduction, and are driven by standardization and waste elimination. Support investments will not generate long term competitive advantage as competitors can easily replicate standardized investments. In cloud computing IaaS provides a low cost replacement infrastructure for the organization's infrastructure.

IV. CASE STUDIES

A number of case studies were developed to examine the Benefits Management approach. The case studies concentrate on how organizations utilize cloud computing, Semi-structured interviews were used to gather information.

A. Organization A – Micro Start-up Company

Organization A was a small start-up organization with less than five employees. It used cloud technology based on PaaS and public cloud technology. The organization provides solutions to the music promotion industry. New artists create music demonstrations that are distributed to Disc Jockeys (DJs), radio stations and music venues. The company has a surprisingly large customer base, with over 2,000 customers and receives several hundred demonstration music tracks each week. A sophisticated feedback and metric system for those listening to demonstration recordings is maintained.

A major operational benefit from using cloud computing is the ability to quickly create new environments for development and live systems that can be easily moved in the cloud.

Music tracks are held in cloud storage, rather than being sent as attachments in e-mails. This gives a number of benefits. There is a single place for the storage of tracks, which has a triple backup. Tracks are downloaded on demand, thus less communications bandwidth is used.

The public cloud can be shared by Organization A and customers. This gives benefits in terms of customer intimacy. Customers can suggest changes to software that can be quickly prototyped and moved rapidly to live systems.

PaaS allows smaller innovative organizations to use the performance enhancements and the unique functionality of cloud computing. The figure below shows the benefits dependency network for Organization A, which shows the enablers, changes and benefits described.

Significant savings have been made by using cloud infrastructure of around £2,400 (UK pounds) per year. The organization has become more agile and can turn around bug fixes and deploy new functionality more quickly.

Quality of service to those receiving music tracks has been improved, as fewer e-mails are rejected due to attachment size. E-mail preparation and delivery was streamlined by using cloud storage. Delivery times were quicker by using high levels of cloud resources for the production of e-mails in short bursts.

Organization A and customers can work together on a shared development platform. This improved communication and reduced configuration problems. The benefits from the benefits dependency network were structured as shown in the table IV.

The structured benefits were used to generate an investment portfolio which is shown in table V.

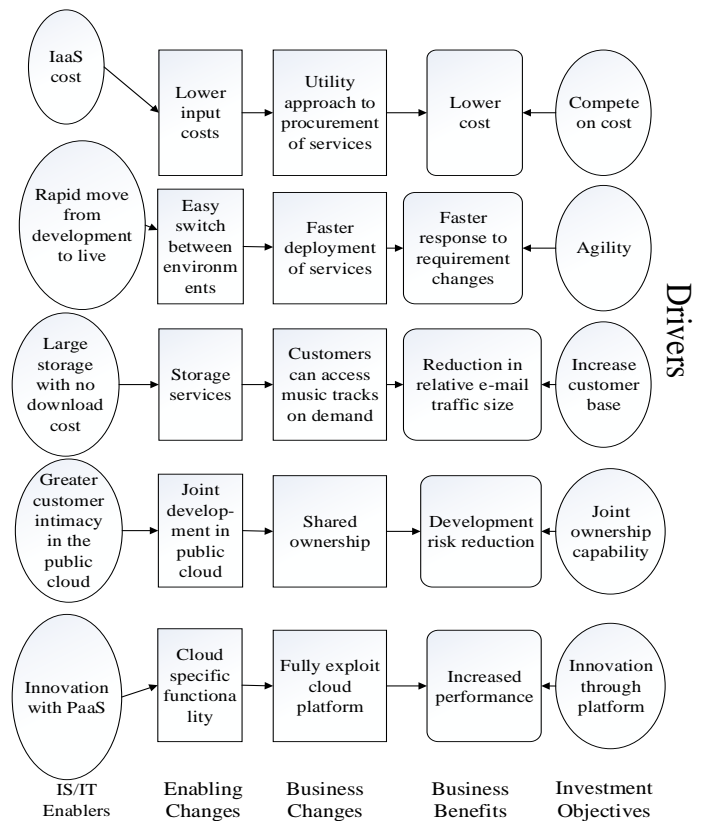


Fig. 2. Benefits Dependency Network for Organization A

TABLE IV. CLASSIFICATION OF BENEFITS FOR ORGANIZATION A

| Degree of Explicitness | Do New Things | Do Things Better | Stop Doing Things |
|------------------------|--|--|---|
| Financial | Lower cost of ownership Around £200 per month | Develop on the cloud, reduces time creating environments and saves around £100 per month | Managing own infrastructure used to cost around £100 per month |
| Quantifiable | Improved quality of service. Problem report to deployment reduced by 2 days | Faster turnaround of new functionality Release cycle cut by 10 person days | |
| Measureable | Less e-mails rejected by music reviewers 13% improvement | Increased performance through PaaS innovations | Sending attachments in e-mails. E-mail send time improved by 19 minutes |
| Observable | Better intimacy with customer | | |

TABLE V. INVESTMENT PORTFOLIO FOR ORGANIZATION A

| | |
|--|-----------------------------------|
| Strategic Investments | High Potential Investments |
| IaaS Cloud storage | PaaS |
| Existing customers with their own hardware | Non-cloud based software |
| Key Operation Investments | Support Investments |

Cloud storage offers a good solution for storing large numbers of high quality music tracks. PaaS offers a mechanism for Organization A to gain competitive advantage over other organizations by using features that are unique to cloud computing, such as the ability to use a large number of virtual processing resources on demand and specialist storage systems.

B. Organization B – Actuarial Services Consultancy

Organization B is an actuarial science consultancy of around 200 employees. The organization has been taken over by a larger organization. Although the organization is now part of a larger organization, they have maintained their independence and adopt an agile management style.

The organizations product is based on building complex economic models, which are used to supply research reports to customers, such as investment banks. Previously, the models were generated on off-the-shelf high-end personal computers or expensive grid-computers. The organization utilized public cloud computing solutions based on IaaS/PaaS platforms. This has allowed economic models to be built more quickly and, at a lower cost than using in-house hardware. The total cost of ownership of infrastructure has also been reduced, as only the resources that are used are paid for and there is no administration or depreciation of IT infrastructure. The figure below shows the BDN for the Organization.

Cost is an enabler for the organization. Prior to cloud resources being available specialist high end personal computers or grid computers were required to run economic scenarios in an acceptable timeframe. The storage of each economic scenario also required several gigabytes of information, with added backup and storage infrastructure costs.

The on-demand resources provided by cloud computing allowed the elimination of waste from the organization. The expensive IT infrastructure previously described wouldn't be renewed when it came to the end of its life, as it was only used 10-20% of the time.

The ability to use potentially unlimited resources on demand was an enabling technology. Economic model scenarios were being run on a quarterly basis for most customers. Larger customers would run scenarios more frequently. The marketing aspects of cloud computing were key enablers in the adoption of cloud computing. Organization B provided a customer solution that was ahead of their competitors, by slightly reworking their well-designed software to function on a PaaS platform.

FIGURE 1 - BENEFITS DEPENDENCY NETWORK FOR ORGANIZATION B

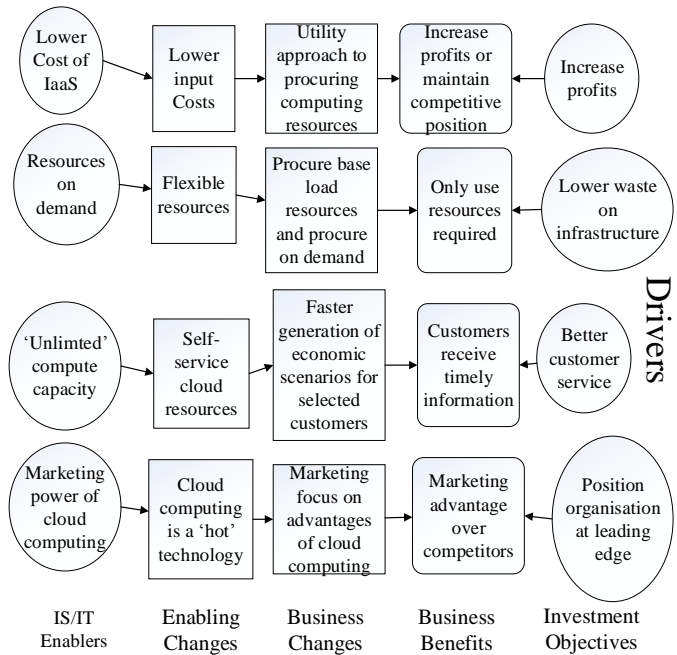


Fig. 3. Benefits Dependency Network for Organization B

The benefits in the BDN were structured, as shown in the table below:

TABLE VI. CLASSIFICATION OF BENEFITS FOR ORGANIZATION B

| Degree of Explicitness | Do New Things | Do Things Better | Stop Doing Things |
|------------------------|---|--|--|
| Financial | Total cost of scenario generation \$12.15 (USD) per scenario | | Grid Computing in the long term. Even grid computers servers cost \$24,900 to \$39,900 (USD) |
| Quantifiable | | Speed up scenario generation by using on-demand processing | |
| Measureable | Reduced additional costs through usage of cloud computing | Saving information in the cloud improves data security and aids customer support | Saving data files and scenarios on customers machines |
| Observable | Improved marketing image | | |

The structured benefits from the investment portfolio are shown below.

TABLE VII. INVESTMENT PORTFOLIO FOR ORGANIZATION B

| Strategic Investments | High Potential Investments |
|--|--------------------------------------|
| IaaS Cloud storage | PaaS Marketing of Cloud Computing |
| Clustered Computer Servers with cloud extensions | Grid Computing |
| Key Operation Investments | Support Investments |

The investment portfolio sees the usage of existing servers being clustered to provide the required processing power to generate economic scenarios. The clusters have been extended to include cloud based processing.

The long term strategy will see a move to cloud based PaaS as in-house servers and grid computers become outdated. Cloud storage was used to store scenario input data and results.

PaaS allows new computer software to be created that generates competitive advantage by using cloud features, rather than running existing software that runs on IaaS. The marketing of software that was capable of running on the cloud generates competitive advantage by being the first or an early adopter in the market.

C. Organization C - Public Sector Division of Large Software Company

Organization C was a large software development organization with 546 staff and £57.2M (UK pounds) in revenues. The organization specialized in a number of markets, such as transport and public sector administration.

The public sector division has been heavily involved in the UK government G-Cloud project. G-Cloud aims to provide a framework for the government provision of cloud computing by companies across IaaS, PaaS and SaaS [23]. Organization C also supplies specialist cloud services to individual customers outside the G-Cloud program.

The benefits dependency network shown above sees the G-Cloud processes and platforms being pivotal to the future of Organization C's public sector division. The core cloud models used are IaaS and SaaS.

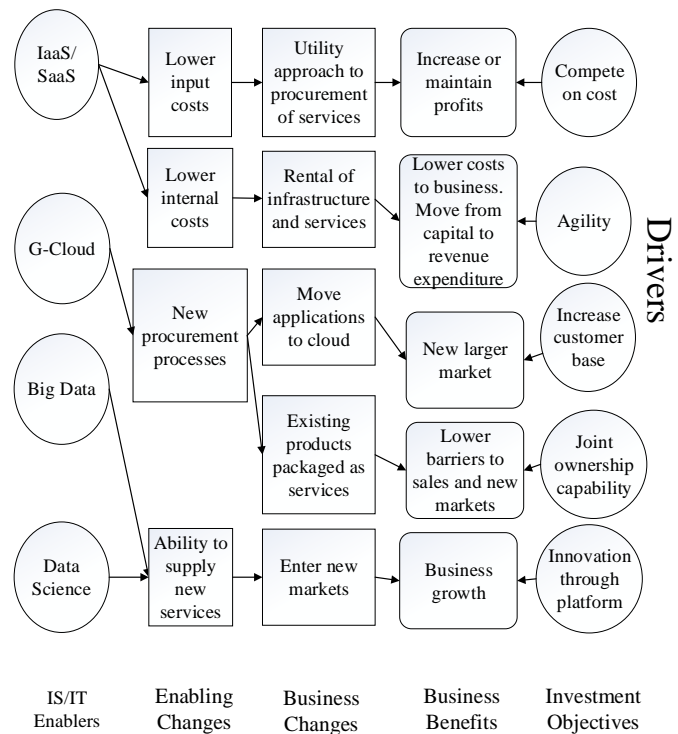


Fig. 4. Benefits Dependency Network for Organization C

New high value and growth sectors are Big Data and Data Science applications built on cloud platforms.

Organization C only uses IaaS and SaaS models and, sees the two services as utilities that can be purchased from a number of suppliers. Cloud computing replaces their current infrastructure, delivery mechanisms to customers of in-house hardware and hosted services. Organization C's internal infrastructure is being replaced by cloud infrastructure to lower costs.

The G-Cloud program and platforms provide access to new markets and change the way software is delivered. New markets have been opened up, allowing greater access to UK government markets and new overseas markets. The cost of entry into markets has been lowered. The organization must be proactive in finding customers.

Emerging Big Data and Data Science enablers has allowed new services to be developed, for example economic analytics, which can be sold to government departments. Benefits are structured in the table below

TABLE VIII. CLASSIFICATION OF BENEFITS FOR ORGANIZATION C

| Degree of Explicitness | Do New Things | Do Things Better | Stop Doing Things |
|------------------------|--|---|---|
| Financial | G-Cloud market potentially worth £11-16 billion. | | |
| Quantifiable | Attack market - 520 times increase in G-cloud market in 2012-2013 | 90% cost saving using SaaS and IaaS for internal infrastructure | Maintaining internal infrastructure to save money |
| Measureable | New Markets for Big Data and Data Science | | |
| Observable | Move from project based solutions to product based solutions, to create utility products | Actively market to customers | Waiting for customers to come to Organization As prime public sector provider |

The table below shows Organization C’s investment portfolio. The enablers from the benefits dependency network create strategic and high potential investments. The main investments are in the G-Cloud program. Existing investments in private and public clouds are supported or are key operational investments in the portfolio.

TABLE IX. INVESTMENT PORTFOLIO FOR ORGANIZATION C

| Strategic Investments | High Potential Investments |
|----------------------------------|-------------------------------------|
| SaaS (utility) IaaS (utility) | G-Cloud Big Data Data Science |
| Private Cloud | Public Cloud |
| Key Operation Investments | Support Investments |

D. Organization D – Public Sector Managed Services

Organization D was an innovative shared service information technology operation between two local government authorities and an IaaS/SaaS provider. Organization D has used cloud computing to provide a school academy finance management system that uses SaaS.

The advantage of this approach was seen as outsourcing some hardware (IaaS) and software (SaaS) ownership to an outside organization. This allowed the expertise of the provider to be used and, therefore reduced risk and the cost of ownership of infrastructure. The disadvantages were a loss of control in the development of the service, control of costs and loss of knowledge within the local authority.

All public sector organizations in the UK operate under the Value for Money (VFM) framework, this forces organizations to consider cost savings which are driving the rapid uptake of

cloud computing. Flexibility is a second enabler for the organization’s adoption of cloud computing. The organization can use their existing hardware for a “base load” and additional resources can be provided by public clouds.

The shared ownership model introduced into the organization with cloud computing allows the organization to share risk with commercial partners and gain commercial income. The organization can also acquire new skills.

Improved service delivery is the final enabler. There was greater availability of infrastructure. Users can use self-service systems to request new software services. New licenses for software are automatically billed back to managers, who could view bills via management information dashboards. The BDN for Organization D is shown in the figure below.

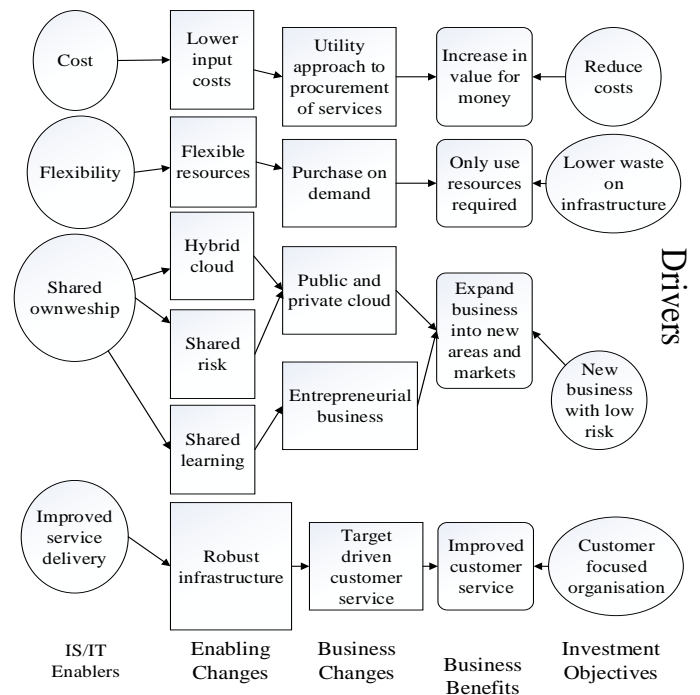


Fig. 5. Benefits Dependency Network for Organization D

The benefits from the BDN are structured in the table below. Major costs savings have been made on infrastructure. The self-service aspects of the cloud infrastructure have reduced the number of service calls. The availability of infrastructure and applications is now approaching 100%. Customer satisfaction has increased from 3.9/5.0 to 4.1/5.0. The local authority has managed to attract two new external customers.

Organization D’s investment portfolio is shown below. The main investments are in IaaS and SaaS. Cloud infrastructure has replaced existing in-house and hosted hardware and software. The organization can bid for new business outside the public authority.

TABLE X. CLASSIFICATION OF BENEFITS FOR ORGANIZATION D

| Degree of Explicitness | Do New Things | Do Things Better | Stop Doing Things |
|------------------------|---|---|---|
| Financial | £9.2M in savings over 5 years | Year on year savings of £1.1 | |
| Quantifiable | 30% reduction in service calls using self service | Availability of ICT systems 99.21% Availability of ICT Infrastructure 99.94% | Reduction of 10 staff servicing desktop computers |
| Measureable | | Customer satisfaction 4.01/5 | |
| Observable | 2 new customers outside councils | | |

TABLE XI. INVESTMENT PORTFOLIO FOR ORGANIZATION D

| Strategic Investments | High Potential Investments |
|----------------------------------|---|
| SaaS (utility) IaaS (utility) | Shared services Grow new business outside councils |
| Private Cloud | Public Cloud |
| Key Operation Investments | Support Investments |

V. RESULTS AND ANALYSIS OF CASE STUDIES

The cloud models are important to how organizations used cloud enablers. This is shown in the table below.

TABLE XII. PROVISION MODELS USED BY ORGANIZATIONS

| Models | Description | Case Study |
|--------|---|---|
| IaaS | Utility based, driven by cost | Organization D Some Organization C Some Organization B Some Organization A |
| SaaS | Can be utility based or used to generate new products | Some Organization D Some Organization C |
| PaaS | Used to generate new concepts and products and greater intimacy with customer | Some Organization B Mostly Organization A |

All of the organizations in the case studies use IaaS directly or as a layer below PaaS/SaaS. IaaS lowers direct costs and reduces management overheads. IaaS is becoming a utility, and is reducing in price that can be purchased by any organization (especially in the public cloud).

The table above shows that Organization A has been able to use PaaS to develop unique products that give it a competitive advantage in the market it operates in. Organization D is relying on IaaS the most, to reduce costs.

SaaS has a dual aspect, as many SaaS offerings are redevelopments of existing CRM and ERP software packages with innovations in pricing and ownership. New innovative SaaS products are also being developed, such as cloud storage,

which is a disruptive technology in the disaster recovery and archival market.

The second dimension of cloud computing is the ownership model, which is shown in the table below.

TABLE XIII. CLOUD OWNERSHIP MODELS OF ORGANIZATIONS

| Ownership | Description | Case Study |
|---------------|----------------------|-----------------------------------|
| Public Clouds | Utility based | Organization A, Organization B |
| Hybrid Clouds | Some utility aspects | Organization D, Organization C |

Public clouds are a utility that will not deliver long term competitive advantage. Organization A has used public clouds as a low cost way to enter new markets and has relied on PaaS to innovate. Organization B is not concerned with developing competitive advantage through ownership models, as it relies on PaaS in the medium to long term. In the short term all organizations will use lower costs and marketing to generate competitive advantage.

Organization C and Organization D can use their size to purchase cloud resources at low cost. Both organizations can also invest in private clouds to develop a unique capability for processing Big Data and Data Science tasks.

Organization D used hybrid cloud technology as a strategic enabler. It combined private cloud infrastructure with low cost public cloud provision. The innovation of seamlessly merging the ownership models and selling IaaS and SaaS products to other organizations provides it with a competitive advantage.

The marketing of cloud computing capability is an important enabler of change and competitive advantage as highlighted in Organization B.

The operational aspects of cloud computing will also provide short term benefits to organizations, especially when using IaaS. These benefits can easily be recreated and will not provide long term competitive advantage.

Cloud models and ownership enablers provide propensities for utility and transformation in organizations. An organization using IaaS in a public cloud is likely to be in a utility market unless other enablers in the BDN coupled with powerful changes are able to deliver benefits. In the case studies a shallow analysis of Organization D would view their heavy reliance on IaaS as a utility approach. However, the usage of hybrid clouds coupled with a strong change culture has delivered benefits and an investment portfolio that has captured new customers. An organization using PaaS and a private cloud has enablers that deliver many benefits. However, without a BDN to deliver the benefits the enablers would be lost.

The power of short term enablers should not be underestimated. Organization B has used marketing as a powerful short term enabler. The leadership understand this can be replicated by competitors in the long term. The organization will then use PaaS as a longer term enabler.

All organizations have used operational gains from cloud computing, to generate competitive advantage in the short term and have strategies in place for the longer term.

VI. CONCLUSION

This paper has described a number of aspects of cloud computing in terms of procurement and ownership models. The Benefits Management approach has been defined. A major contribution of the paper is to undertake a number of case studies on diverse organizations to understand how these organizations use procurement and ownership models to derive benefits from cloud computing.

The conclusions from the case studies outline the beginnings of a “database” which can be used by organizations when considering cloud computing investments. Further case studies will add to the knowledge and robustness of the portfolio of case studies.

Another contribution of this paper is to consider the expressive power of existing Benefits Management tools. It is felt that the tools may not be powerful enough to consider the complexity of cloud computing investments across a number of organizations. Work is underway to consider more powerful knowledge representation techniques.

VII. FUTURE WORK

On reviewing the case studies presented in this paper it can be seen that there is a complexity in and between cloud enablers, change, benefits and investments. The tools provided by traditional Benefits Management provide basic network interconnections between elements and simplistic representations of the Benefits Management elements.

A more powerful knowledge representation method is required. It is proposed to model the Benefits Management process as an ontology and to populate the ontology with the case studies described in this paper and further case studies.

The Benefits Management ontology is strongly related to pricing models already developed for cloud computing [24]. The two ontologies can be mapped to consider the value of benefits derived from cloud computing investments.

REFERENCES

- [1] J. Ward and E. Daniel, *Benefits Management: How to Increase the Business Value of Your IT Projects*. Wiley, 2012.
- [2] M. Cusumano, “Cloud computing and SaaS as new computing platforms,” *Communications of the ACM*, vol. 53, no. 4, pp. 27–29, 2010.
- [3] R. L. Grossman, “The Case for Cloud Computing,” *IT Professional*, vol. 11, no. 2, pp. 23–27, Apr. 2009.
- [4] P. Mell and T. Grance, “The NIST definition of cloud computing (draft),” *NIST special publication*, vol. 800, no. 145, p. 7, 2011.
- [5] D. Agrawal, “Towards the End-to-End Design for Big Data Management in the Cloud: Why, How, and When?,” presented at the BTW, 2013, pp. 15–16.
- [6] I. T. Foster and R. K. Madduri, “Science as a service: how on-demand computing can accelerate discovery,” presented at the Proceedings of the 4th ACM workshop on Scientific cloud computing, 2013, pp. 1–2.
- [7] R. Moreno-Vozmediano, R. S. Montero, and I. M. Llorente, “Key Challenges in Cloud Computing: Enabling the Future Internet of Services,” *Internet Computing, IEEE*, vol. 17, no. 4, pp. 18–25, 2013.
- [8] C. Low, Y. Chen, and M. Wu, “Understanding the determinants of cloud computing adoption,” *Industrial management & data systems*, vol. 111, no. 7, pp. 1006–1023, 2011.
- [9] S. C. Misra and A. Mondal, “Identification of a company’s suitability for the adoption of cloud computing and modelling its corresponding Return on Investment,” *Mathematical and Computer Modelling*, vol. 53, no. 3, pp. 504–521, 2011.
- [10] Y. Han, “Cloud computing: case studies and total cost of ownership,” *Information technology and libraries*, vol. 30, no. 4, pp. 198–206, 2011.
- [11] A. B. Mohammed, J. Altmann, and J. Hwang, “Cloud computing value chains: Understanding businesses and value creation in the cloud,” in *Economic models and algorithms for distributed systems*, Springer, 2010, pp. 187–208.
- [12] M. E. Porter, “Towards a dynamic theory of strategy,” *Strategic Management Journal*, vol. 12, no. S2, pp. 95–117, 1991.
- [13] J. Peppard, J. Ward, and E. Daniel, “Managing the realization of business benefits from IT investments,” *MIS Quarterly Executive*, vol. 6, no. 1, pp. 1–11, 2007.
- [14] J. Peppard and A. Rylander, “From Value Chain to Value Network:: Insights for Mobile Operators,” *European Management Journal*, vol. 24, no. 2–3, pp. 128–141, Apr. 2006.
- [15] J. Ward and E. Daniel, “How to deliver more business benefits from IT investments,” *The European Financial Review*, pp. 27–30, 2013.
- [16] J. S. Harrison, D. A. Bosse, and R. A. Phillips, “Managing for stakeholders, stakeholder utility functions, and competitive advantage,” *Strategic Management Journal*, vol. 31, no. 1, pp. 58–74, 2010.
- [17] J. Highsmith, *Agile Project Management: Creating Innovative Products*, 2nd ed. Addison-Wesley Professional, 2009.
- [18] E. Daniel, “Breakfast Briefing presentation 29th November 2012.” Open University, 2012.
- [19] H. Wehrich, “The TOWS matrix—a tool for situational analysis,” *Long range planning*, vol. 15, no. 2, pp. 54–66, 1982.
- [20] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and others, “A view of cloud computing,” *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.
- [21] D. C. Wyld, “The Cloudy future of government IT: Cloud computing and the public sector around the world,” *International Journal of Web & Semantic Technology*, vol. 1, no. 1, pp. 1–20, 2010.
- [22] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, “Big data, analytics and the path from insights to value,” *MIT Sloan Management Review*, vol. 52, no. 2, pp. 21–31, 2011.
- [23] M. R. Catherine and E. B. Edwin, “A Survey on Recent Trends in Cloud Computing and its Application for Multimedia,” *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 2, no. 1, p. pp–304, 2013.
- [24] R. Greenwell, X. Liu, and K. Chalmers, “Pricing Intelligence as a service for Cloud computing.” *IEEE Cloudcom 2013*, 2013.

Image Segmentation Via Color Clustering

Kaveh Heidary

Department of Electrical Engineering and Computer Science
Alabama A&M University
4900 Meridian Street, Huntsville, AL 35810 USA

Abstract—This paper develops a computationally efficient process for segmentation of color images. The input image is partitioned into a set of output images in accordance to color characteristics of various image regions. The algorithm is based on random sampling of the input image and fuzzy clustering of the training data followed by crisp classification of the input image. The user prescribes the number of randomly selected pixels comprising the trainer set and the number of color classes characterizing the image compartments. The algorithm developed here constitutes an effective preprocessing technique with various applications in machine vision systems. Spectral segmentation of the sensor image can potentially lead to enhanced performance of the object detection, classification, recognition, authentication and tracking modules of the autonomous vision system.

Keywords—Clustering; Classification; Image Segmentation; Machine Vision

I. INTRODUCTION

Background removal and image segmentation constitute fundamental components of many autonomous vision systems. Segmentation is utilized in order to separate regions or entities of potential interest from each other and from inconsequential image background for further processing. This paper presents an operationally robust and computationally efficient algorithm for segmentation of the input image based on color. The complex image at the sensor output, which is presented to the autonomous vision system, is partitioned into multiple less complicated images prior to further processing [1-5]. Following the segmentation phase, pertinent members of the resultant image set are processed by the corresponding target classification, recognition, identification, and authentication layers of the machine vision system.

The image segmentation process at lower levels entails ascribing to each pixel the appropriate class label, while at the higher levels segmentation involves utilization of lower level information for associating salient parts of the image with known objects of interest [6-10]. Target detection, classification, recognition, and authentication procedures which are based on two dimensional spatial signatures acquired with various modalities including infrared and electro optical imagery involve utilization of spatial filters [11-15]. The spatial filters may be applied directly to the image at the sensor output or to the resultant images following the segmentation stage. This paper provides the formulation and implementation of an efficient lower level image segmentation algorithm. Image pixels are classified in accordance to their color attributes regardless of spatial relationships. The color classifier is computed using a set of

randomly selected pixels, obtained from the input image, which are partitioned using a fuzzy clustering procedure. A set of prototype color vectors are computed from the resultant fuzzy sets and are subsequently utilized to segment the input image.

II. BACKGROUND

Image segmentation is used in order to partition the input image into its salient components for further processing. Segmentation is utilized in various machine vision applications such as object recognition and tracking as well as image compression, editing, and retrieval. Segmentation involves clustering the image feature vectors such as pixel intensity levels and colors [16-18]. In top-down image segmentation the input image is partitioned in accordance to the relationship between the image content and the images of various objects in the database including object shapes, contours, textures, and colors. The bottom-up image segmentation, on the other hand, utilizes the intensity, color, texture, and region boundaries to break up the image into its more basic components. Despite the impressive results of recently reported bottom-up image based segmentation algorithms, they often fail to capture fundamental relationships among image elements. The inherent difficulty encountered by low-level image based segmentation algorithms is due to potentially sharp intensity and color variations within the object boundaries. High-level segmentation algorithms rely on image features such as contours and shapes as segmentation primitives in order to reduce the computational complexity. Detection of edges and contours in the input image is achieved through convolving the grayscale image with local derivative filter operators [19-20]. Different regions that are circumscribed by distinct closed contours are subsequently recognized as the respective image segments. This Paper presents an unsupervised learning algorithm for segmentation of color images.

III. CLUSTERING ALGORITHM

Given a set of N data points in M -dimensional space, and a user specified integer representing the number of clusters (classes) Q , the algorithm described here computes a set of Q prototypes and a $Q \times N$ membership matrix. Each prototype is a vector in M -space and is the optimal representation of the corresponding class. Each element of the membership matrix represents the degree of membership (association) of a data point in the respective cluster.

$$\mathbf{X} = \mathbf{x}_n : 1 \leq n \leq N \quad (1)$$

$$\mathbf{X}_n = x_{nm} : 1 \leq m \leq M \quad ; 1 \leq n \leq N \quad (2)$$

$$\mathbf{Y} = \mathbf{Y}_q : 1 \leq q \leq Q \quad (3)$$

$$\mathbf{Y}_q = y_{mq} : 1 \leq m \leq M ; 1 \leq q \leq Q \quad (4)$$

$$x_{mn}, y_{mq} \in \mathbb{R} \quad (5)$$

Where \mathbf{X}, \mathbf{Y} represent, respectively, the set of data points and prototype vectors in M -space, and \mathbb{R} is the set of real numbers. Our objective is to utilize data points in Eq. (1) in order to partition M -space into Q distinct regions with each region represented by a prototype vector \mathbf{Y}_q . In the operation phase, an unlabeled vector is classified based on its distance with respect to the prototypes. In crisp classification, for example, the input vector is assigned uniquely to the class with the closest prototype with respect to the input vector. In fuzzy classification, on the other hand, the input vector is assigned to all classes with varying degrees of association.

Each original data point in the trainer set will be linked to all Q regions (classes) with varying degrees of association determined by elements of the membership matrix. The initial membership matrix is generated by assigning random numbers drawn from independent and identically distributed uniform probability functions to each matrix element. We will describe an iterative algorithm for computation of the prototype vectors. The prototype vectors are then used to make hard decisions with regard to new input data points. A new data point is associated with the prototype (class) to which it is closest in accordance to some predefined distance metric.

$$\mathbf{S} = s_{qn} ; 1 \leq q \leq Q, 1 \leq n \leq N \quad (6)$$

$$s_{qn} = \frac{\alpha_n}{\sum_{p=1}^Q d_{qn} / d_{pn}^{\frac{2}{u-1}}} \quad (7)$$

$$\alpha_n = \frac{1}{\sum_{q=1}^Q \frac{1}{\sum_{p=1}^Q d_{qn} / d_{pn}^{\frac{2}{u-1}}}} \quad (8)$$

$$d_{pn} = \|\mathbf{Y}_p - \mathbf{X}_n\| \quad (9)$$

Where, \mathbf{S} is the membership (association) matrix, s_{qn} denotes the degree with which data point- n is associated with (is member of) cluster- q , and d_{qn} is the distance between data point- n and prototype- q . Here, Euclidean distance is used as a measure of distance between vectors in M -space. The exponent parameter $u \in [1, \infty)$ is user-specified and determines the fuzziness of the clustering process. It is noted from Eq. (8) that the membership matrix is normalized such that sum of each column is equal to one. When $u = \infty$, each data point belongs to all clusters uniformly and $s_{qn} = 1/Q, 1 \leq n \leq N, 1 \leq q \leq Q$. when $u = 1$, however, clustering is not fuzzy and each data point is associated with a unique cluster. For crisp (hard) clustering, elements of the membership matrix are given as follows:
 $s_{qn} = 1, d_{qn} < d_{pn} \forall p \neq q$ and $s_{qn} = 0$ otherwise. In hard

clustering, $u = 1$, each column of \mathbf{S} contains a single one and the rest of entries for that column are zero. The value of u affects the rate of convergence of the algorithm. In experiments conducted on diverse sets of RGB images, we have found that setting $u = 2.5$, in general, leads to fast convergence and accurate results.

The process starts with generating a random membership matrix, called the zero-order membership matrix $\mathbf{S}^{(0)}$. Matrix elements are chosen from a uniform probability distribution function $s_{qn}^{(0)} \in [0, 1]$. The matrix is then normalized by setting the sum of each column to one. The randomly generated membership matrix is then utilized to compute Q zero-order prototype vectors, one for each cluster. A particular prototype vector is computed as the weighted sum of the entire set of data points, where each data point is weighted in accordance to its association to (membership in) the respective cluster.

$$\mathbf{Y}^{(0)} = \mathbf{Y}_q^{(0)} : 1 \leq q \leq Q \quad (10)$$

$$\mathbf{Y}_q^{(0)} = \frac{\sum_{n=1}^N s_{qn}^{(0)u} \mathbf{X}_n}{\sum_{n=1}^N s_{qn}^{(0)u}} ; 1 \leq q \leq Q \quad (11)$$

Where, $\mathbf{Y}_q^{(0)}$ represents the zero-order prototype vector associated with cluster- q , \mathbf{X}_n is the n th data vector denoting a typical trainer, $s_{qn}^{(0)} (1 \leq q \leq Q, 1 \leq n \leq N)$ are elements of the randomly generated zero-order membership matrix, and u is the user-specified exponential parameter. The zero-order prototype vectors are then utilized to compute the first-order membership matrix as shown below.

$$\mathbf{S}^{(1)} = s_{qn}^{(1)} ; 1 \leq q \leq Q, 1 \leq n \leq N \quad (12)$$

$$s_{qn}^{(1)} = \frac{\alpha_n^{(0)}}{\sum_{p=1}^Q d_{qn}^{(0)} / d_{pn}^{(0)\frac{2}{u-1}}} \quad (13)$$

$$\alpha_n^{(0)} = \frac{1}{\sum_{q=1}^Q \frac{1}{\sum_{p=1}^Q d_{qn}^{(0)} / d_{pn}^{(0)\frac{2}{u-1}}}} \quad (14)$$

$$d_{pn}^{(0)} = \|\mathbf{Y}_p^{(0)} - \mathbf{X}_n\| \quad (15)$$

$$\mathbf{G}^{(1)} = \mathbf{g}_{qn}^{(1)} ; \mathbf{g}_{qn}^{(1)} = |s_{qn}^{(1)} - s_{qn}^{(0)}| \quad (16)$$

$$\delta^{(1)} = \max_{q,n} \mathbf{g}_{qn}^{(1)} \quad (17)$$

Where, $\mathbf{S}^{(1)}$ and $\mathbf{G}^{(1)}$ denote, respectively, the first order membership and gradient matrices, and $\delta^{(1)}$ is the first order gradient. Next, the computed first order membership matrix is used in order to compute the first order prototype vectors using Eq. (11), where the superscript 0 is replaced with 1. subsequently, the computed first order prototype vectors are

utilized to compute the second order membership matrix and the gradient as shown in Eq. (13) and Eq. (17), respectively.

The iterative process described above continues until a user prescribed stopping criterion is met. The stopping criterion may be the maximum number of iterations (orders), in which case the process is terminated when the number of iterations is reached. One may also use the gradient value or the relative change of gradient between two consecutive iterations as the stopping criterion. For the experiments in this paper the iteration process terminates when the gradient falls below the user prescribed threshold, i.e. $\delta^{(r)} < T = 0.001$.

IV. TESTS WITH SIMULATED DATA

The clustering algorithm described above was used to partition various synthetically generated data sets into classes, numbers of which were prescribed by the user. The algorithm computes a prototype for each class and the membership matrix for the entire data set. Each trainer may then be assigned to a unique class by binarizing the computed membership matrix. Likewise, new unlabeled input data are classified based on distance between the data point and the computed prototypes.

In the example of Figure 1, the input data set is comprised of points in the xy-plane of the Cartesian coordinate system. the data points were generated by a pair of 2D Gaussian distributions with means at (1,3), (-2,-1) and equal standard deviations set to one. The xy components of each data point were generated by independent distributions. Fifty points were randomly selected from each distribution and were combined to form the unlabeled set of one-hundred input data points. Parameters of the clustering algorithm were set to $Q = 2, u = 2, T = 0.001$, and the iteration process converged after ten rounds. Figure 1 shows the evolution of two prototypes. Both prototype vectors started very close to each other at the proximity of the center of gravity of the entire data set. As the iterations proceed, it is seen that prototypes move toward centers of the respective distributions, where final values of the computed prototypes are shown as triangles.

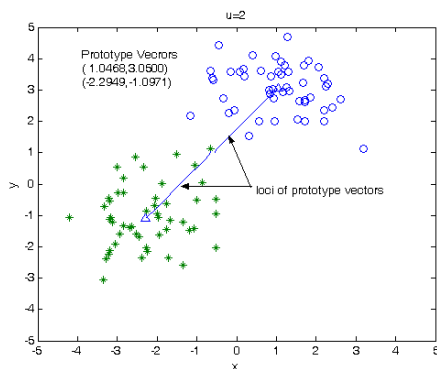


Fig. 1. Evolution of prototype vectors in a two-class problem.

In Fig 1, using circles and stars to denote the data points associated with two classes is for illustration purpose only, and the algorithm is entirely oblivious to the class dispositions of input data points. Despite complete lack of knowledge about origins of the data set members in Figure 1, it is seen

that the algorithm finds appropriate prototypes for two classes. Figure 2 shows the degree of association of various data points (1-100) to each prototype (class), and is an illustration of the computed membership matrix for the fuzzy classifier. It is seen that the first fifty data points are more strongly associated with the first prototype ($q=1$), whereas the last fifty points have higher association to the second prototype ($q=2$), as expected. Fuzzy classification may be utilized for assignment of classes to new unlabeled input data, where each input vector is assigned probabilities of membership in respective classes. In some applications, crisp classification of the input data may be desired, where a typical input vector is assigned exclusively to the class whose prototype is closest, based on Euclidian distance, to the input vector.

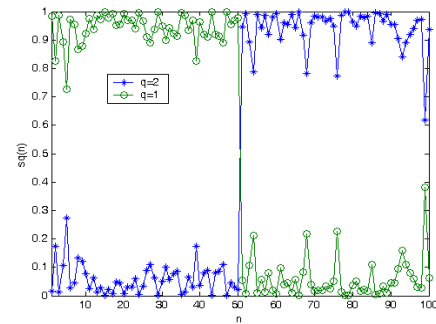


Fig. 2. Trainer data association factor.

In the example of Figure 3, the input data set comprises 125 points in the xy-plane, randomly chosen from three Gaussian distributions with means at (1,3), (-2,-6), (3,-2), and different variances along two axes. The clustering algorithm was tasked to partition the above unlabeled set of data using parameters $Q = 3, u = 2, T = 0.001$. As expected, all three prototype vectors are initially very close to each other and are situated virtually at the center of gravity of the entire input data set. As the iteration process proceeds the prototypes traverse the xy-plane toward their final destinations denoted as triangles. It is noted that the computed prototypes in this example are not equal to mean vectors of the respective classes. This is a byproduct of dataset composition and does not affect ability of the computed prototype vectors to accurately classify new and unlabeled input data.

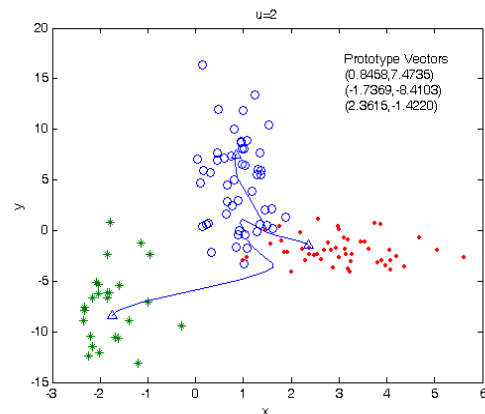


Fig. 3. Prototype evolutions in a three-class problem

In order to obtain quantitative performance results for the fuzzy clustering algorithm the following tests were conducted. We started with two Gaussian distributions in 2D-space, where the x and y components of each sample point were obtained from independent distributions with equal variances.

In the first example, classes A and B were obtained from two 2D distributions with equal variances $\sigma_A = \sigma_B$. Equal numbers of randomly generated trainers from each class were combined and were subsequently utilized as the unlabeled trainer set.

Fuzzy clustering was applied to the above set of unlabeled trainers in order to evolve two prototypes. A large number of test points were then generated from the distribution functions described above. Binary classification was utilized to label all the test vectors in accordance to their Euclidean distances with respect to the above two computed prototypes.

Figure 4 shows the classification error rate as function of the separation factor with the number of unlabeled trainers utilized from each class as parameter.

$$SF = \frac{\|m_A - m_B\|}{\sqrt{\sigma_A^2 + \sigma_B^2}} \quad (18)$$

Where, SF represents the separation factor between two distributions, and m, σ denote, respectively, the mean-vector and the standard deviation of the particular data set. For each test case, the number of trainers was fixed and the separation factor was varied from 0.25 to 4.

Error rate is the percentage of input test vectors that are misclassified. As expected, the classifier performance improves as the separation factor increases. It is noted that number of trainers has virtually no effect on classifier performance.

In the example of Figure 5, the two Gaussian distributions have unequal standard deviations such that $\sigma_B = 2\sigma_A$ and all other parameters are same as before. It is seen that the number of trainers in this case has a slightly more pronounced effect of the classifier performance.

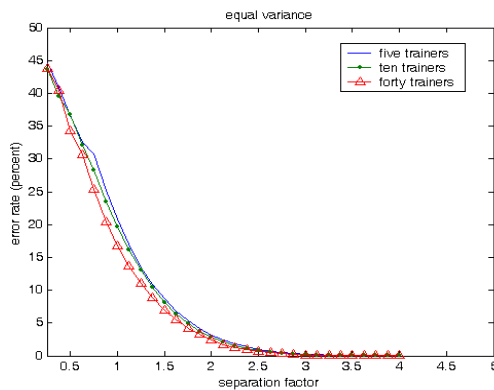


Fig. 4. Effect of SF on classification error rate

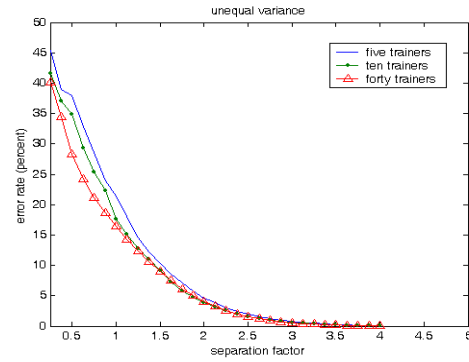


Fig. 5. Effect of SF on classification error rate

V. EXPERIMENTS WITH REAL DATA

In this section the fuzzy clustering algorithm described above is applied to the task of segmentation of color images. The set of RGB vectors associated with a group of randomly selected pixels of the input color image constitute the trainer set. The computed prototypes comprise a set of color vectors which are subsequently utilized to partition the input image.

The example of Figure 6 shows the input image (upper-left), and the result of color segmentation. In this example one-hundred pixels ($N=100$) were randomly selected from the input image, comprising the unlabeled trainers which were used as the input of the fuzzy clustering process. The algorithm was tasked to partition the training set into three classes ($Q=3$). Figures 7 and 8 show, respectively, the trainers and the evolution of class prototypes in the RGB-space. It is noted that all three prototypes are initially very close to each other and are proximate to centroid of the training set. The prototypes migrate toward their factual positions and true prototypes are evolved as shown in Figure 8. The initial and final values of the RGB coordinates of the prototype vectors for three classes are listed in Table 1. In this example it took twenty iterations for all three prototypes to reach their final destinations. The computed prototypes were then utilized for crisp classification of all the input image pixels. One of three possible labels were assigned to each pixel of the input image. The images of Figure 6 show results of the filtering process.



Fig. 6. Original input image (upper-left), and color-segmented images. Classes-one (upper right), two (lower-left), and three (lower-right)

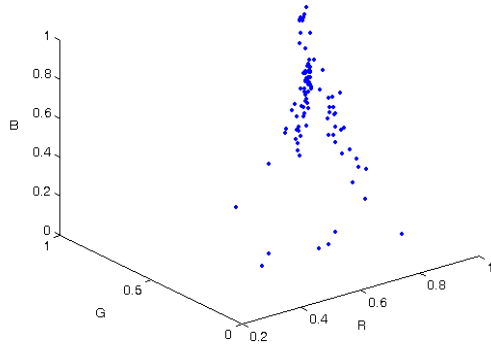


Fig. 7. Training set comprised of one-hundred randomly selected pixels from the original input image

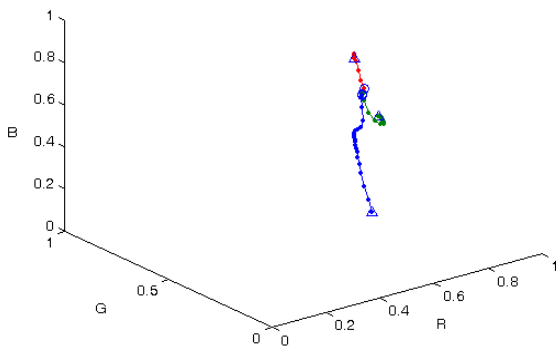


Fig. 8. Evolutions of prototypes for classes-one through three. Initial estimates of prototypes are denoted as circles and final estimates are triangles.

TABLE I. INITIAL AND FINAL PROTOTYPE VECTORS.

| | Initial Prototypes | | | Final Prototypes | | |
|-------------|--------------------|-------|-------|------------------|-------|-------|
| | R | G | B | R | G | B |
| Class-one | 0.853 | 0.678 | 0.492 | 0.566 | 0.255 | 0.233 |
| Class-two | 0.861 | 0.687 | 0.492 | 0.946 | 0.716 | 0.345 |
| Class-three | 0.880 | 0.700 | 0.507 | 0.885 | 0.757 | 0.623 |

In the next example the trainer set consisted of one-hundred pixels randomly selected from the Mondrian painting of Figure 9. Fuzzy clustering was used to partition the trainer set into four classes, and the respective RGB prototype vectors were computed.

All pixels of the input image were subsequently classified crisply in accordance to the prototype with the smallest Euclidean distance with respect to the corresponding pixels. The images of Figure 10 show the result of input image segmentation, where pixels of the corresponding class are turned on while pixels of all other classes are set to white.

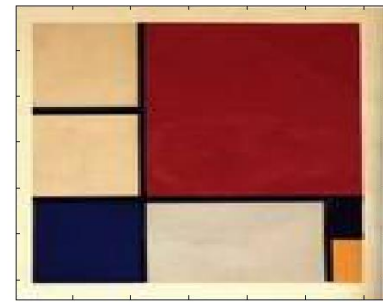


Fig. 9. Input image

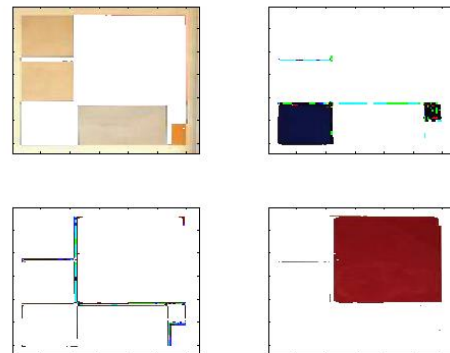


Fig. 10. Output images.

Figures 11 and 12 show the set of one-hundred training vectors and the evolution of four prototype vectors for the Mondrian of Figure 9. As before, all four prototypes are initially close to the center of mass of the training set. In this experiment, The process converged after six iterations. Circles and triangles in Figure 12 denote, respectively, the initial and final values of the prototype vectors. The above four computed prototype RGB vectors were used to assign each pixel of the input Mondrian to one exclusive class, characterized by the prototype with the smallest Euclidean distance with respect to the RGB vector of the pixel. Each one of the filtered images in Figure 10 shows the pixels of the respective class with all other pixels set to white.

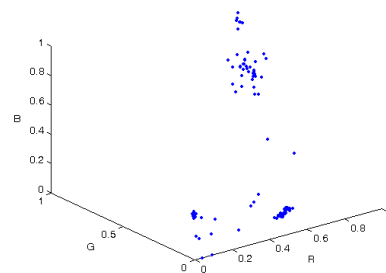


Fig. 11. Training set comprised of one-hundred randomly selected pixels of the Mondrian

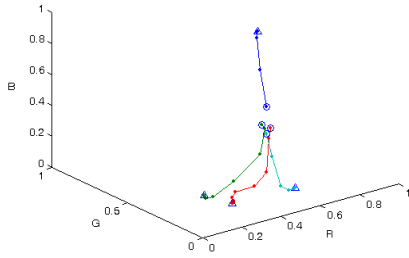


Fig. 12. Evolutions of prototypes for classes-one through four. Initial estimates of prototypes are denoted as circles and final estimates are triangles.

The images of Figure 13 show an input image and the resultant filtered images which are produced by the prototype vectors computed from fuzzy clustering of the trainer set into three groups. The one-hundred element trainer set, shown in Figure 14, was obtained by random sampling of the input image. Figure 15 shows the evolution of the prototype vectors.



Fig. 13. Upper left shows the input image. The input image is filtered using a three-class filter.

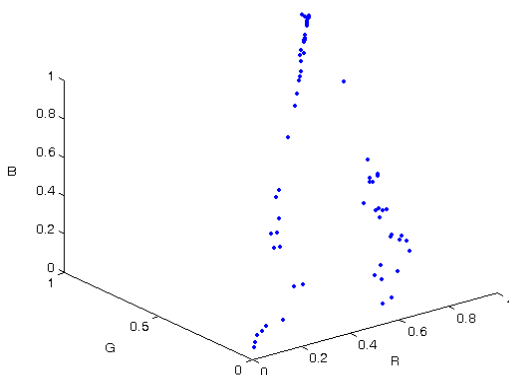


Fig. 14. Training set comprised of one-hundred randomly selected pixels of the input image.

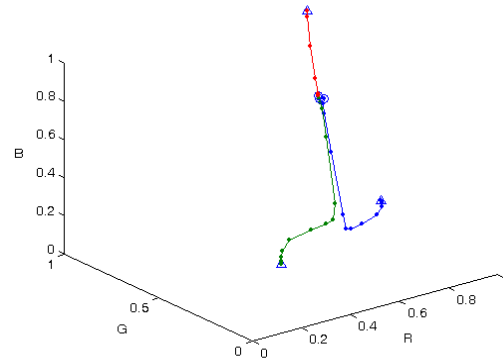


Fig. 15. Evolutions of prototypes for classes-one through three. Triangles denote final prototypes.

In the next example the input image is partitioned in a hierarchical manner. First, the image is sampled randomly and the samples are grouped into two classes using the fuzzy clustering algorithm. This leads to computation of two prototypes, which are used to carry out crisp segmentation of the input image. This process produces two images, each comprised of the input image pixels that belong to the respective class with all other pixels set to white. Each of the two generated images is treated as a new input image and is partitioned into two classes, resulting in four new images. The process continues for a user specified number of partition rounds.

The original set of training pixels were selected randomly from the input image of Figure 16, and were partitioned into two classes using fuzzy clustering. The images of Figure 17 show the result of this two-class segmentation process. The Class-one image, consisting of the leaf and bug only which constitute the foreground in the original image, was then sampled randomly to form a new set of trainers which were partitioned using fuzzy clustering, resulting in computation of two new prototypes. The image (Class-one) was subsequently partitioned using the computed prototypes. The images of Figure 18 show the segmentation results.

VI. CONCLUSIONS

This paper provides a computationally efficient and operationally robust algorithm for segmentation of color images. Tests using synthetically generated data sets as well as real RGB images have demonstrated the efficacy of the image segmentation procedure developed here. The algorithm has practical applications in machine vision systems where partitioning the sensor images in accordance to color characteristics of various image regions can precede higher level processing layers such as recognition and tracking of targets. Future work will include utilization of different distance measures such as Mahalanobis distance and applications to multi-spectral image segmentation.



Fig. 16. Input image.

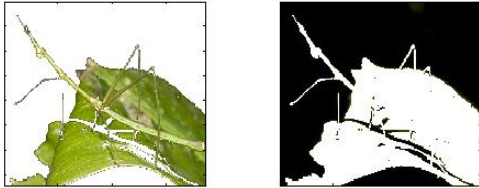


Fig. 17. The result of segmentation of the input image into two classes, foreground and background.

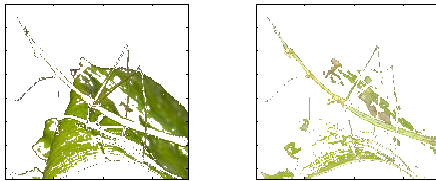


Fig. 18. The result of segmentation of the foreground into two classes

VII. ACKNOWLEDGEMENTS

Partial sponsorship for this work was provided by Department of Defense through US Army RDECOM W911NF-13-1-0136.

REFERENCES

- [1] Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** (8), 2002, pp. 1026–1038.
- [2] Cheng, H.D., Jiang, X.H., Sun, Y., Wang, J.: Color image segmentation: advances and prospects. *Pattern Recog.* **34** (12), 2001, pp. 2259–2281.

- [3] Heidary, K., Caulfield, J. :Color Classification using margin-setting with ellipsoids. *Signal Image and Video Processing*. 2012, DOI 10.1007/s11760-012-0349-6.
- [4] Heidary, K., Caulfield, J. : Presmoothing effects in Artificial Color image segmentation. *Computer Vision and Image Understanding* **117**, 2013, pp. 195-201.
- [5] Heidary, K., Caulfield, J. :Discrimination among similar looking noisy color patches using Margin Setting. *Optics Express* **15** (1), 2007, pp. 62-75.
- [6] Batchelor, B.G. (Editor), *Machine Vision Handbook*, Springer, 2012.
- [7] Steger, C., Ulrich, M., Wiedemann, C.: *Machine Vision Algorithms and Applications*, Wiley -VCH, 2007.
- [8] Davies, E.R.: *Computer and Machine Vision: Theory, Algorithms, Practicalities*, Academic Press, 2012.
- [9] Acharya, T., Ray, A.K.: *Image Processing - Principles and Applications*, Wiley, 2006.
- [10] Zhu, H., Zheng, J., Cai, J., Thalman, N.M.: Object-level image segmentation using low level cues. *IEEE Transactions on Image Processing* **22** (10) 2013, pp. 4019-4027.
- [11] Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24** (24) 2002, pp. 509-522.
- [12] Bhanu, B.: Automatic target recognition: State of the art survey. *IEEE Transactions on Aerospace and Electronic Systems* **22** (4) 1986, pp. 364-379.
- [13] Heidary, K.: Distortion tolerant correlation filter design. *Applied Optics* **52** (12) 2013, pp. 2570-2576.
- [14] Heidary, K.: Synthetic template: effective tool for target classification and machine vision. *International Journal of Advanced Computer Science and Applications* **4** (10) 2013, pp. 22-31.
- [15] Heidary, K., Caulfield, H.J.: Needles in a haystack: fast spatial search for targets in similar-looking backgrounds. *Journal of the Franklin Institute* **349** 2012, pp. 2935-2955.
- [16] Martin, D., Fowlkes, C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26** (5) 2004, pp.530–549.
- [17] Yu, S., Shi, J.: Multiclass spectral clustering. *Proceedings of International Conference on Computer Vision*, 2003.
- [18] Lezoray, O., Charrier, C.: Color image segmentation using morphological clustering and fusion with automatic scale selection. *Pattern Recognition Letters* **30**, pp. 397-406, 2009.
- [19] Marr, D.C., Hildreth, E.: Theory of edge detection. *Proceedings of Royal Society of London*, 1980.
- [20] Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and Hierarchical Image Segmentation. *IEEE Trans. Pattern Analysis Machine Intelligence* **33** (5) 2011 pp. 898-916.

A Framework to Improve Communication and Reliability Between Cloud Consumer and Provider in the Cloud

Vivek Sridhar

Rational Software Group (India Software Labs)
IBM India
Bangalore, India

Abstract—Cloud services consumers demand reliable methods for choosing appropriate cloud service provider for their requirements. Number of cloud consumer is increasing day by day and so cloud providers, hence requirement for a common platform for interacting between cloud provider and cloud consumer is also on the raise. This paper introduces Cloud Providers Market Platform Dashboard. This will act as not only just cloud provider discoverability but also provide timely report to consumer on cloud service usage and provide new requirement based on consumer cloud usage and cost for the same. Dashboard is also responsible for getting cost of each service provider for a particular requirement. Our solution will learn from requirements and provide required details for consumer for effective usage of cloud services. This also enable service provider to understand requirements, provide quality of service, to understand new requirement and deliver. This framework also deals with how best we can use before and after usage of cloud services to choose a right service provider for a particular requirement in a community.

Keywords—cloud computing; requirement communication; requirement engineering; cloud service; cloud discoverability; data Mining; artificial intelligence in Cloud Computing

I. INTRODUCTION

Some analysts and vendors define cloud computing narrowly as an updated version of utility computing: basically virtual servers available over the internet. Cloud computing comes into focus only when you think about what IT always needs: a way to increase capacity or add capabilities on the fly without investing in new infrastructure, training new personnel, or licensing new software [7].

There are set of services and models working behind the scene making the cloud computing feasible and accessible to end users. Following are the models for cloud computing:

- 1) *Deployment models*
- 2) *Services Models*

Deployment models define the type of access to the cloud, i.e., how the cloud is located? Cloud can have any of the four types of access: public, private, hybrid and community. Service models are the reference models on which the cloud computing is based. There are three basic categories:

- 1) *Infrastructure as a Service (IaaS)*
- 2) *Platform as a Service (PaaS)*

3) *Software as a Service (SaaS)*

A cloud provider is a company that offers some component of cloud computing – typically above three services to other businesses or individuals. Cloud providers are referred to as cloud service providers (CSP). There are a number of things to think about when you evaluate cloud providers. The cost usually be based on per-user utility model but there are a number of variations to consider. Reliability is very essential if your data must be accessible. A typical cloud infrastructure service-level agreement (SLA), for example, specifies explicit levels of service – such as, for example, 99.9% uptime – and the compensation that the user is eligible to if the provider fails to provide the service as described [8]. Hence it becomes difficult for both cloud provider and cloud consumer to match each other's requirements without a common platform.

Cloud services are seen as a deployment model, enabling users to consume software and hardware services via internet. These services are supplied by various services providers. Cloud services are designed to provide easy, scalable access to applications, resources and services, flexible pricing and customization. Cloud computing will soon become a utility [1] available anytime, anywhere.

The rest of the paper is structured as follows. Section II presents the research problem. In Section III, introduce cloud providers market platform dashboard as solution to the research problems defined in section II.

II. RESEARCH PROBLEM

With the development of cloud computing systems, consumers' requirement also become increasingly complex, this leads to following five problems between consumers and providers.

1) *Due to diverse cloud services providers, consumers very often find it difficult to choose right services for their requirement. Also they don't have information on other consumer's chosen service provider for similar kind of requirement. For example, if cloud consumer has requirement for virtual servers having - operating system IBM AIX 7.1 / RAM 10 GB with network bandwidth 100 Mbps then consumer has no information on which cloud provider is offering best service for this particular requirement. No information on which service provider to choose for this particular*

requirement. Also there is no method to find out which service provider did the other consumers choose for similar requirement.

2) Consumers are not aware of provider's reliability / usability / performance / scalability / interoperability for the specific requirements. There is no method or algorithm to find quality of service provided by the provider for a particular requirement. Consumer need to understand his application requirement and then choose appropriate provider. Quality of service should be based on consumer requirements.

3) Consumers don't get report on cloud usage and hence don't get recommendation on whether to continue the service or to upgrade the service with same provider or to choose new provider. Once the consumer starts consuming cloud services there is no proper channel to understand consumer usage and get recommendation.

4) Consumers don't have information whether to switch cloud provider for cost effective cloud usage.

5) Cloud service providers demand new architecture to understand requirement coming from such heterogeneous cloud consumers. Also providers don't get information on consumers chosen provider for the same requirement and reasoning for the same. This will help provider to analyze their requirement engineering [6] strategy and improve quality of service to expand their offerings.

Hence in this paper we are introducing a framework which attempts to solve all the problems noted above and help consumer to get better cloud service for their applications. Most of the on going research deals with stakeholders functions and does not address more specific topics as noted above. Some research work deals with discoverability of cloud services but doesn't deal with recommendations before discoverability. None of the research work deal with how best we can use before and after usage of cloud service to choose right service for a particular requirement in a community.

III. CLOUD PROVIDERS MARKET PLATFORM DASHBOARD

Our primary research goal is to provide a dashboard to consumer to key in there requirements and get recommendations to choose cloud service provider. The same tool will be used by provider to understand requirements, analyze and expand offering. This dashboard uses StakeCloud Platform concept [2]. StakeCloud Platform takes in requirement input and search for cloud services provider who provides such service and match the results. It doesn't take our research problems into consideration. It only deals with cloud service discoverability and doesn't provider recommendations through learning.

Cloud providers market platform dashboard deals with how best we can use before and after usage of cloud service to choose the right provider for a particular requirement in a community. This platform requires providers to open there service discoverability and cloud monitoring features. In the course of development provider may have to open various other features for the consumer benefit.

To meet the research goal, we depicted Cloud Providers Market Platform Dashboard in Fig 1. The framework contains 8 components. Dashboard is a community based where all consumers and all providers are using the same platform for communication. All providers who wish to be part of this dashboard should adhere to standards defined by Engine component.

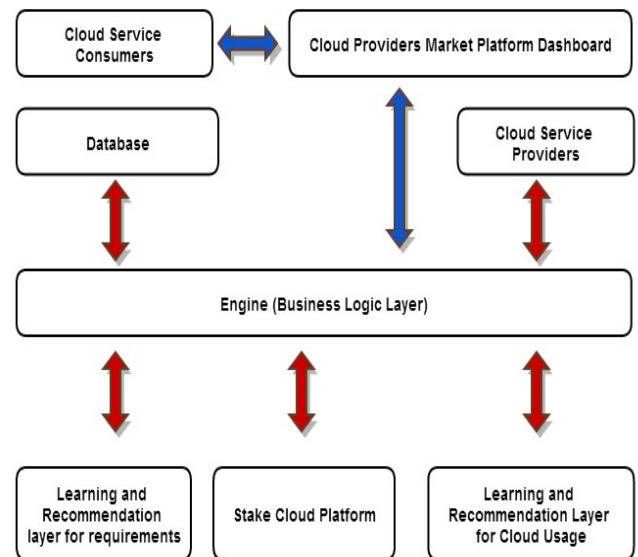


Fig. 1. Cloud Providers Market Platform Dashboard

“ABC Inc” is a company which needs virtual server on cloud. Consider “ABC Inc” as cloud consumer. Scenario based demonstration is done using all 8 components of the framework defined above.

Step by Step scenario to illustrate the framework

- Cloud consumer (ABC Inc) access “Cloud Provider Market Platform Dashboard” via HTTP / HTTPS protocols and access requirement request form to input his requirement.
- Consumer inputs the requirement in the dynamic web platform. Consumer gets to choose his requirement for his application from the web. For Example: Which server operating system, Disk size, RAM details, Bandwidth details, IPV6 support, required software's (like database / IDE) etc. Consumer waits for the result once he submit job to Dashboard as in Fig.2.
- Once consumer submits his requirement, dashboard inputs data into Database component through Engine component (Could be any of the database on the cloud for example: Informix / MySQL or Oracle) and pass the control to Engine via internet protocols. Engine component acts as a master and is sole responsible for driving this framework. All the communication between Engine and to other component is via HTTP / HTTPS protocol.

Server Operating System

Please select Server Operating System.

Disk Size

RAM

Bandwidth

IP Version
 IPv6 Support
 IPv4 Support

Required Software
 Select Software's required on the server

Rational Synergy
 Python
 Java
 MySQL Database
 Informix Database
 Rational Team Concert
 Rational functional Tester
 Oracle Database
 Perl
 VMWare API's
 Eclipse Client Version 4.2

From Date * / /

End Date * / /

Fig. 2. Requirement Request Form

- The requirement submitted by cloud service consumer is processed by 2 layers through Engine: 1. Learning and Recommendation layer for requirements 2. Stake Cloud Platform [2].
- Learning and Recommendation layer has the ability to collect data from Engine layer and pass information to dashboard based on pervious matching requirements from various consumers and their selected cloud providers details. Also usage report of the same. This component learns every requirement coming in and keeps track of all the activity related to requirement. This helps the consumer to choose right cloud service provider. This layer also has the ability to provide information on provider's reliability / usability / performance / scalability etc from Cloud service usage layer Fig.5.

Consumer requirement

Server Operating System : AIX 7.0
 Disk Size : 200 GB
 RAM : 10 GB
 Bandwidth : 100 Mbps
 IP Version : IPv6 Support , IPv4 Support
 Required Software : Rational Team Concert , Python , Java
 From Date : 07 / 09 / 2014
 End Date : 08 / 07 / 2014

Other Consumers chosen provider for the same requirement

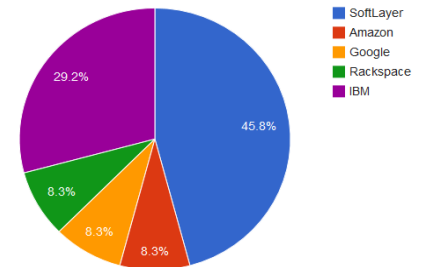


Fig. 3. Other consumers of the same requirement and there provider details (overall chart)

- Stake Cloud Platform [2] has been enhanced to provide quotation (cost) for the services from various providers after the match is found for the requirement. Requirements are processed and list the matching result in the dashboard Fig 4. Above 2 layers are responsible for listing cloud service providers for the consumer requirement.

Cloud Provider list from Stake Cloud Platform Component with cost(for a particular requirement)

| Cloud Service Provider | Requirements | Cost | Cost Effective |
|------------------------|--|---------------|----------------|
| Softlayer | Server Operating System : AIX 7.0 Disk Size : 200 GB RAM : 10 GB Bandwidth : 100 Mbps IP Version : IPv6 Support , IPv4 Support Required Software : Rational Team Concert , Python , Java | \$200 per day | Yes |
| IBM | Server Operating System : AIX 7.0 Disk Size : 200 GB RAM : 10 GB Bandwidth : 100 Mbps IP Version : IPv6 Support , IPv4 Support Required Software : Rational Team Concert , Python , Java | \$203 per day | Yes |
| Amazon | Server Operating System : AIX 7.0 Disk Size : 200 GB RAM : 10 GB Bandwidth : 100 Mbps IP Version : IPv6 Support , IPv4 Support Required Software : Rational Team Concert , Python , Java | \$308 per day | No |
| Google | Server Operating System : AIX 7.0 Disk Size : 200 GB RAM : 10 GB Bandwidth : 100 Mbps IP Version : IPv6 Support , IPv4 Support Required Software : Rational Team Concert , Python , Java | \$310 per day | No |
| Rackspace | Server Operating System : AIX 7.0 Disk Size : 200 GB RAM : 10 GB Bandwidth : 100 Mbps IP Version : IPv6 Support , IPv4 Support Required Software : Rational Team Concert , Python , Java From Date : 07 / 09 / 2014 End Date : 08 / 07 / 2014 | \$307 per day | No |

Fig. 4. Cloud provider list from stake cloud platform component with cost (for a particular requirement)

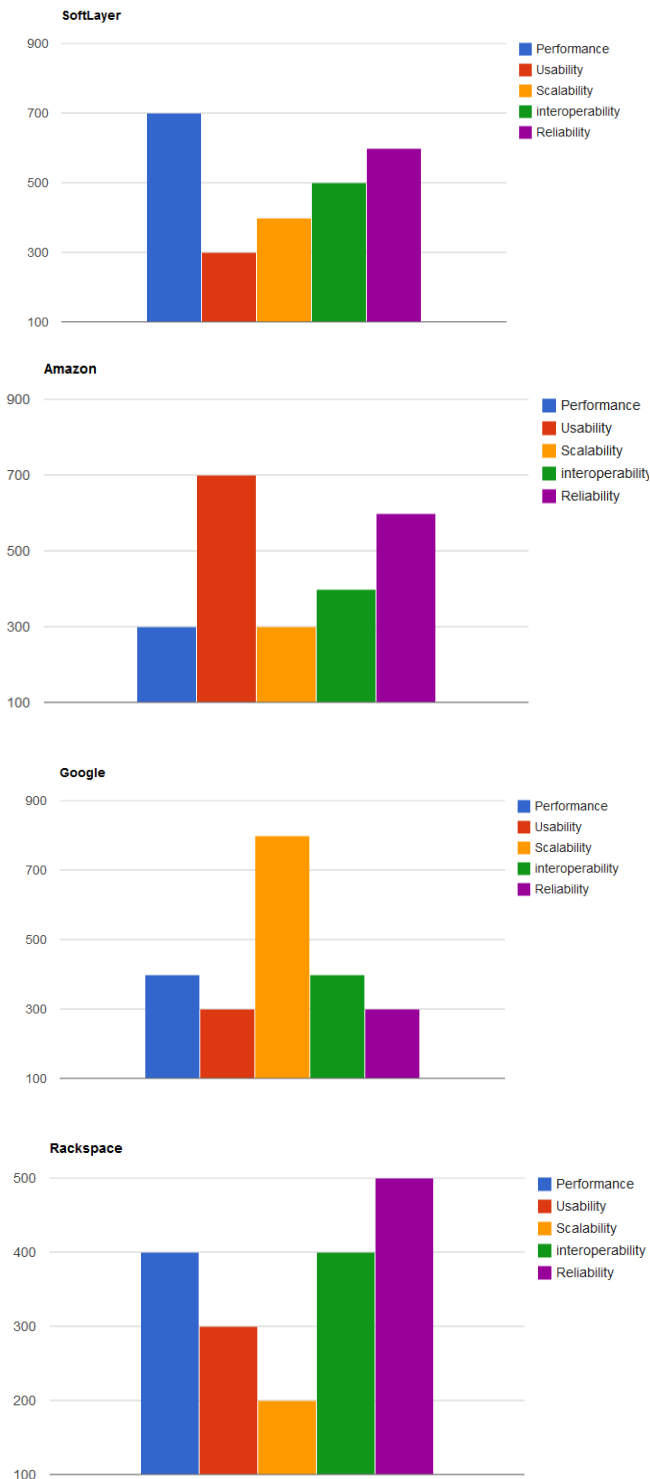


Fig. 5. Cloud providers Quality of Service for the requirement in question

- Once the consumer selects a services provider and start's consuming the service, the learning and recommendation layer for cloud usage is activated. This layer has the ability to store consumers chosen service

provider and also monitor its activity (as part of SLA [5]). Note that for these activities cloud service provider is actively involved and interact with Engine component. This layer provides timely report to consumer through dashboard on cloud service usage. Also responsible for upgrade or down-grade SLA's [5] with recommendation of switching between different providers for cost effective cloud usage. This report is always based on cloud usage for a particular application and its expense as seen in Fig.6.

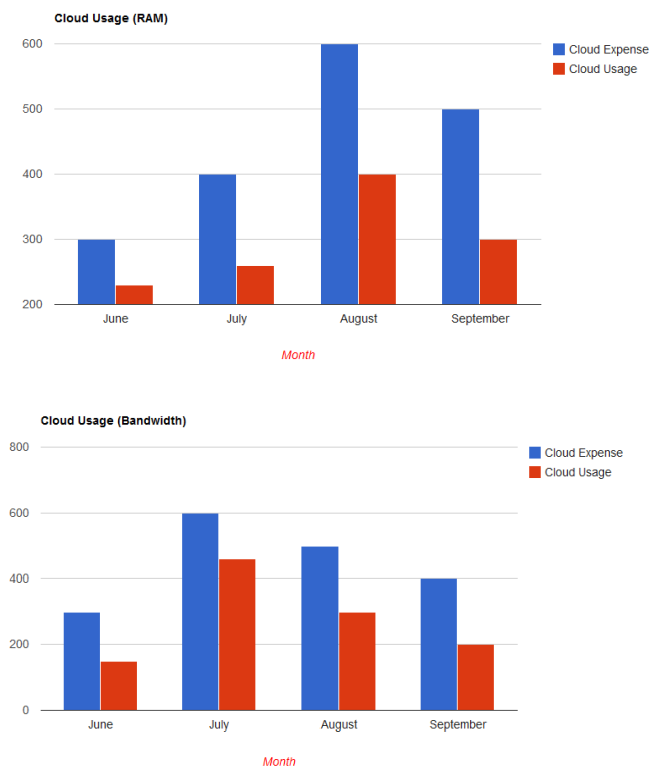


Fig. 6. Cloud Usage report

- Cloud usage report in Fig.6 indicates consumer need to down-grade service level agreement with the provider for cost effective cloud service usage. Also indicates whether consumer needs to switch to different provider for effective usage of cloud services as seen in Fig.7.

| Recommended provider list after cloud usage report | | | |
|--|---|---------------|----------------|
| Cloud Service Provider | Requirements | Cost | Cost Effective |
| Softlayer | Server Operating System : AIX 7.0 Disk Size : 200 GB RAM : 10 GB Bandwidth : 100 Mbps IP Version : IPv6 Support , IPv4 Support Required Software : Rational Team Concert , Python , Java | \$200 per day | Yes |
| IBM | Server Operating System : AIX 7.0 Disk Size : 200 GB RAM : 10 GB Bandwidth : 100 Mbps IP Version : IPv6 Support , IPv4 Support Required Software : Rational Team Concert , Python , Java | \$320 per day | No |
| Amazon | Server Operating System : AIX 7.0 Disk Size : 200 GB RAM : 10 GB Bandwidth : 100 Mbps IP Version : IPv6 Support , IPv4 Support Required Software : Rational Team Concert , Python , Java | \$210 per day | Yes |
| Google | Server Operating System : AIX 7.0 Disk Size : 200 GB RAM : 10 GB Bandwidth : 100 Mbps IP Version : IPv6 Support , IPv4 Support Required Software : Rational Team Concert , Python , Java | \$310 per day | No |

Fig. 7. Recommended provider list after cloud usage report

- All the data generated are stored in Database component for data retrieval at later stage. Experiments

carried out are part of stimulation and doesn't involve real time data.

- Cloud service provider learns all the activity through Engine to improve their requirement engineering strategy and improve quality of service to expand their offerings. Note that cloud service provider component is dynamic (always in sync with Engine layer) and constantly updated, which is a drawback of Stake Cloud Platform [2] where service offerings are extracted from provider. This might not be true at the time of consuming service since its not dynamic Fig 8.

Consumers chosen chart for particular requirement

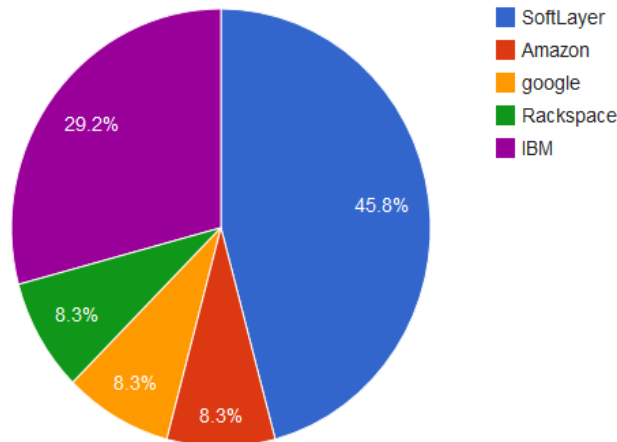


Fig. 8. Consumers Chosen chart for particular requirement.

Cloud provider market platform dashboard is driven by data. As data is collected, Engine layer writes new rules to retrieve data for consumer to choose right service provider for that particular requirement. Also provider understand requirement and enhance their requirement strategy to provide matching quality of service.

IV. FUTURE WORK AND CONCLUSION

The main contributions of this research lie in enabling consumers to find the best mapping cloud services for their requirement, and in supporting providers to identify real consumer needs. Above research experiment is part of stimulation and doesn't involve real time data.

Future work includes implementing this research work involving all the stake holders in real time. And also to improve performance and automatic switching between providers for cost effective cloud usage. There is a need to integrate SLA trust model with the above proposed dashboard [9].

ACKNOWLEDGMENT

This work is carried out in part under Computer Science and Engineering Department of R. N.S Institute of Technology and IBM Rational Software Group (India Software Labs).

REFERENCES

- [1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic. Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25(6):599–616, 2009.
- [2] Todoran, I., & Glinz, M. (2012). Towards bridging the communication gap between consumers and providers in the cloud. *Proceedings of the WICSA/ECSA 2012 Companion Volume on - WICSA/ECSA '12*, 78. doi:10.1145/2361999.2362012.
- [3] S. Lichtenstein, L. Nguyen, and A. Hunter. Issues in it service-oriented requirements engineering. *AJIS*, 13(1):176–191, 2005.
- [4] P. Patel, A. Ranabahu, and A. Sheth. Service level agreement in cloud computing. In *Cloud Workshops at OOPSLA09*, pages 1–10, 2009.
- [5] J. J. M. Trienekens, J. J. Bouman, and M. V. D. Zwan. Specification of service level agreements: Problems, principles and practices. *Software Quality Journal*, 12:43–57, 2004.
- [6] B. Verlaine, I. Jureta, and S. Faulkner. Towards conceptual foundations of requirements engineering for services. In *Proc. RCIS 2011: 5th Intl. Conf. on Research Challenges in Information Science*, pages 1–11. Gosier, 2011
- [7] <http://www.infoworld.com/d/cloud-computing/what-cloud-computing-really-means-031>
- [8] <http://searchcloudprovider.techtarget.com/definition/cloud-provider>
- [9] Alhamad, M., Dillon, T., & Chang, E. (2010). SLA-Based Trust Model for Cloud Computing. 2010 13th International Conference on Network-Based Information Systems, 321–324. doi:10.1109/NBIS.2010.6

A Wavelet-Based Approach for Ultrasound Image Restoration

Mohammed Tarek GadAllah¹ and Samir Mohammed Badawy²

Abstract — Ultrasound's images are generally affected by speckle noise which is mainly due to the scattering phenomenon's coherent nature. Speckle filtration is accompanied with loss of diagnostic features. In this paper a modest new trial introduced to remove speckles while keeping the fine features of the tissue under diagnosis by enhancing image's edges; *via Curvelet denoising and Wavelet based image fusion*. Performance evaluation of our work is done by four quantitative measures: the *peak signal to noise ratio* (PSNR), the *square root of the mean square of error* (RMSE), a *universal image quality index* (Q), and the *Pratt's figure of merit* (FOM) as a quantitative measure for edge preservation. Plus Canny edge map which is extracted as a qualitative measure of edge preservation. The measurements of the proposed approach assured its qualitative and quantitative success into *image denoising while maintaining edges as possible*. A Gray phantom is designed to test our proposed enhancement method. The phantom results assure the *success and applicability of this paper approach not only to this research works but also for gray scale diagnostic scans' images including ultrasound's B-scans*.

Keywords — *Ultrasound Medical Imaging; Curvelet Based Image Denoising; Wavelet Based Image Fusion*.

I. INTRODUCTION

Ultrasound medical imaging which has been widely accepted as an essential safe tool for biological tissue medical diagnosis, are generally affected by speckle noise due to the scattering phenomenon's coherent nature. Speckle noise is a well known phenomenon inherent most B-mode ultrasonic scans' images caused by the constructive and destructive interferences of the wavelets scattered by the tissue components as they arrive at the transducer [1], [2]. Speckle degrades the resolution and contrast of ultrasound images [3]. Speckle noise poses a well known problem in ultrasound imaging [4]. It acts as a mask of the small differences in grey level images [5]. Therefore the pre filtering process of Speckle noise cannot be avoided. It is a critical pre-processing step, providing clinicians with enhanced diagnostic ability [13]. The filtration is accompanied with loss of diagnostic features. The amount of these losses differs according to the techniques reported so far.

¹ **Mohammed Tarek GadAllah**; is with a company; as a Maintenance & Operating Engineer, Broadcast Division, at; Filka Television Broadcasting Station - Filka Island – Kuwait. (E-mail: mohammed.tag.1986@gmail.com), (Phone: 00965-97-466-049)

² **Samir Mohammed Badawy**; is a Doctor Emeritus with the Faculty of Electronic Engineering, Menofia University, Menofia, Egypt. (E-mail: drsamirb@gmail.com), (Phone: 002-0100-8409-506)

For example; the Multiscale Method introduced by *Achim, A. ... et al*; in [29]. The wavelet based denoising is one of the effective filtration techniques, like the technique introduced by *Pizurica, A. ... et al*; in [30], and the method introduced by *Rabbani, H. ... et al*; in [31]. Wavelets have been widely used in signal and image processing for the past 20 years [18], [19], [27], and [28]. Wavelet transform and its derivatives have many applications in biomedical image processing [6]-[13] and [17]-[28]. The first introducing for wavelets into biomedical imaging research was in 1991; by *Weaver, J. B. ... et al*, in a journal paper [21] describing the application of wavelet transforms for noise reduction in MRI images, [19]. One of its derivatives is the *Curvelet Transform* (CVT) which first mentioned by E. J. Candès and D. L. Donoho in 1999 [15]. The *Digital Curvelet Transform* (DCT) was introduced by D. L. Donoho & M. R. Duncan in November 1999 [16]. J.L. Starck, E.J. Candes, and D.L. Donoho published: "The Curvelet Transform for Image denoising" in 2002; [12]. Image denoising in Curvelet domain has enhanced denoising; due to the ability of *Curvelet Transform* to recover signals in different directions as compared with other methods [6]-[13]. Denoising in Curvelet domain has better results for speckle noise reduction of ultrasonic scans' images, but; in some cases *it cannot maintain all features of the scan's image*. This is a well-known problem in the field of biomedical imaging and image processing [32]-[40].

In this paper; we introduce an approach to remove speckles while keeping the fine features of the tissue under diagnosis as possible by enhancing image's edges; *via the Curvelet denoising and Wavelet based image fusion*. The study was done on a normal kidney ultrasonic scan taken from a man; 27 years old, by an ultrasound console: Aloka-ProSound 3500SX, in DICOM format [49]. Beside; an *original Gray phantom* is built to prove the success of our proposed enhancement method into better speckle reduction with edge enhancement for ultrasound scans, than the only Curvelet based denoising; qualitatively and quantitatively. A general quality optimization index S&M (S. Badawy and M. GadAllah) is newly introduced for selecting the best parameter's value for any parameter-based image fusion method being firstly introduced into image processing research. The paper is organized as follows: In Section II; we display the materials and methods had been used in our paper including our proposed approach, been applied on a human right kidney scan and the gray phantom study. The numerical and graphical results are displayed in Section III. A brief discussion of our results is represented in Section IV. Finally, we give some concluding remarks in Section V.

II. MATERIALS AND METHODS

A. Curvelet Based Image Denoising:

The Curvelets is to represent a curve as a superposition of functions of various lengths and widths [11]. Curvelet transforms gave close and improved delineation to edges [14]. Curvelet construction based on three essential ideas: firstly, *Ridgelets*; a method of analysis suitable for objects with discontinuities across straight lines. Secondly, *Multiscale Ridgelets*, a pyramid of windowed Ridgelets, renormalized and transported to a wide range of scales and locations. Thirdly, *Band-pass Filtering*, a method of separating an object out into a series of disjoint scales [15]. We used a Matlab-based Toolbox: [33] made by Sandeep P.; for performing *Image denoising in Curvelet domain using thresholding*. This toolbox [33]; when computing the Curvelet of an image, it makes that as given in [12].

B. Wavelet Based Image Fusion:

The principle of image fusion using wavelets is to merge the wavelet decompositions of the two original images using fusion methods applied to approximations coefficients and details coefficients [50]. There are more techniques for image fusion [41]-[44].

We used the *Wavelet toolbox, built-in Matlab-R2011b* to perform: *wavelet based image fusion*; a kind of a Multiscale-Decomposition Based Fusion. The fusion been made, was in-between the noisy image (N) and the reconstructed (denoised) image (R) produced from the denoising process in Curvelet transform domain. The resulted fused image, we called the final processed image (FP). An illustrative block diagram of the used fusion method is shown in Fig.1, where;

- N is referred to the Noisy image -input.
- R is referred to the Reconstructed image -input.
- A is referred to Approximation.
- D is referred to Details.
- FP is referred to Final Processed image -output.

We applied the *User-Defined Fusion method*; mentioned in [50] for merging the approximation coefficients of the two input images; A_N and A_R , as the following *Matlab code*:

```
function AFP = my_fusion (AN,AR)
D = true (size (AN)); F = 0.50;
AFP = AN;
AFP(D) = F * AN(D) + (1-F) * AR(D);
AFP(~D) = F * AR(~D) + (1-F) * AN(~D);
end
```

The parameter **F**; we called the *fusion ratio*; for the Approximation fusion method. The value of **F** can be changed from **0** to **1**. In our study we used different values of **F**. The best used value of **F** was **0.5**; achieving a maximum image's edge preservation as well as maintaining its quality as possible. So, we recommend using a value of **0.5** for **F**.

C. Performance Evaluation:

1) **RMSE** is the square root of **MSE** calculated from the following equation (1):

$$MSE = \frac{1}{m \times n} \sum_{i=1}^m \sum_{j=1}^n [\|O(i, j) - R(i, j)\|]^2 \quad \dots (1)$$

Where; **m** = number of rows in the image, **n** = number of columns, **O** is the original image pixel values matrix, **R** is the reconstructed image pixel values matrix, and finally **(i,j)** are the **x, y** coordinates for every pixel in each image. **MSE** has been the dominant quantitative performance metric in the field of signal processing for more than 50 years [45]. We had programmed; equation (1) of MSE, on *Matlab-R2011b* and implemented it in our measurements.

2) **PSNR** is represented in the following equation (2):

$$PSNR = 10 \log\left(\frac{MAX_o^2}{MSE}\right) \quad \dots (2)$$

Where, MAX_o is the maximum pixel value in image O [here; $MAX_o = 255$], MSE here is the mean square error calculated from equation (1). We used *Matlab-R2011b* and [33] for calculating PSNR. PSNR and RMSE measures are used in more papers which are about sonograms' denoising [11].

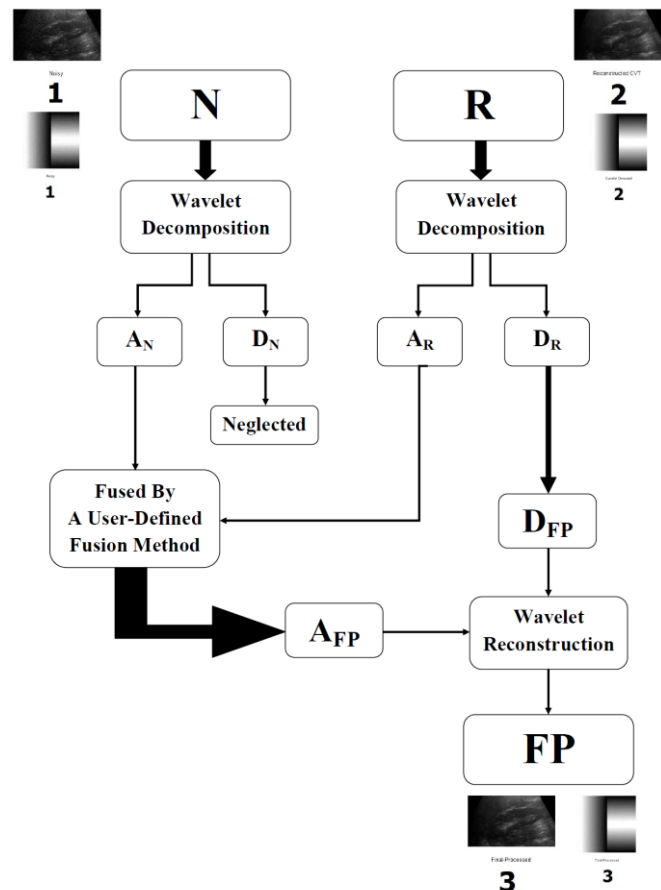


Fig. 1. Illustrative Block Diagram of the Used Fusion Method

3) **Q** is the Universal Image Quality Index, introduced in [46], it is based on three basic components measured in-between the original image and the image which to be compared; correlation coefficient, mean luminance, and contrast [46]. Its value ranges from 0 till 1, if $Q = 1$, it means the two images are same. We used [46] and Matlab to calculate Q in our measurements.

4) **FOM** is the Pratt's Figure of Merit; introduced in 1978 by W. K. Pratt as a quantitative measure for edge preservation [47]. FOM is defined by:

$$R = \frac{1}{I_N} \sum_{i=1}^{I_A} \frac{1}{1+aD^2} \dots (3)$$

Where; $I_N = \text{MAX} \{I_l, I_A\}$ and I_l and I_A represent the number of ideal and actual edge map points; respectively, a is a scaling constant, and D is the separation distance of an actual edge point normal to a line of ideal edge points. The rating factor is normalized so that $R = 1$ for a perfectly detected edge [47]. We used [48] by Matlab to calculate FOM in our measurements, but in percent % values, so if FOM = 100, it means that FOM = 1 in (3).

D. The Proposed Approach

An ultrasound scan of a normal human right kidney was taken by ultrasound console: *Aloka-ProSound 3500SX*; [49]. The interested area of the resulted image had been put on a black background with an appropriate dimensions match with Curvelet transform domain; as shown in Fig.2.

Manipulation was done as follow:

- i. Noising the image by an added noise determined by the value of the Signal to Noise Ratio (**Snr**).
- ii. Denoising the resulted image in Curvelet transform domain, as mentioned previously in **A**.
- iii. Applying the wavelet based fusion method been explained previously in **B**.; in-between the noised and the denoised image produced from **i.** and **ii.**, respectively taking $F = 0.95959595$.
- iv. *Measurements*:
 - The edge detection map of Canny had been extracted as a qualitative measure of edge preservation for each image of the three images produced from **i.**, **ii.**, and **iii.**
 - A four quantitative performance evaluation measures PSNR, RMSE, Q, and FOM; had been calculated in-between the original image and each one of its three derivatives produced in **i.**, **ii.**, and **iii.**
- v. Repeating the previous four steps **i.**, **ii.**, **iii.**, and **iv.**; along **20** different values for Snr.
- vi. The FOM's **20** results is plotted for each image vs. the corresponding Snr, into Fig.3-A, where: the three colored lines; Blue, Green, and Red represents: in the three cases: Noisy, denoised (CVT), and the final wavelet fused image (WT), respectively.

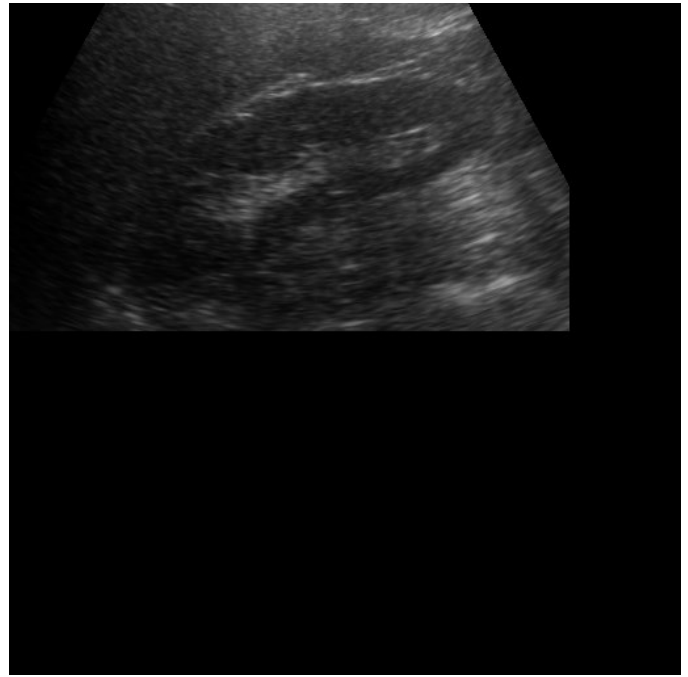


Fig. 2. The Right Kidney Scan On A Black Background

Optimizing the Applied Fusion Method:

Our goal is to find the better performance of the applied fusion method that makes maximum edge preservation as well as maintains image's quality and PSNR as possible. For that we introduce here, a general quality optimization index be named S&M index. This index can be used in general for selecting the best parameter's value (*F in our study*) for any parameter-based image fusion method being firstly introduced into image processing research. S&M index is defined as follow:

$$S \& M = \frac{\%PSNR_n \times \%FOM_n \times \%Q_n}{\%RMSE_n \times 100} \% \dots (4)$$

Where, **_n**: means after normalization to the nearest maximum value. This equation calculates maximum values for PSNR, FOM, Q, and minimum value for RMSE.

We applied S&M optimization for the applied fusion method as follow:

- 1) Taking one result from the right kidney scan's image, where Snr = 10, after step **ii**; i.e. after noising and denoising process, then, we made step **iii** for this result not only at (F) = 0.95959595, But, instead we made this step **25** times for different values of **F**; starting from **0.05** to **1.0**.
- 2) Calculating the four measurements; PSNR, RMSE, Q, and FOM; for all the 25 results, and normalized all to the nearest maximum value.
- 3) Drawing in percentage %; the 25 final wavelet fused numerical results for PSNR, RMSE, Q, and FOM, after being approximately normalized; all vs. the fusion ratio (**F**); as

shown in **Fig.3-B**; where: the four colored lines; Blue, Green, Red, and Blue-Light represents: the four measurements: PSNR, RMSE, Q, and FOM; respectively.

4) Finally, we applied equation (4), for S&M optimization index, on the 25 results from the previous step 2, producing the optimization figure; shown in **Fig.3-C**; where: the y-axis; represents: S&M index value, and the x-axis; represents: **F**; starting from **0.05** to **1.0**.

Repeating Manipulation taking $F = 0.5$:

From equation (4); we can deduce that the optimum performance of the proposed approach, *making maximum edge preservation as well as maintaining image's quality and PSNR as possible*; can be founded where S&M index has a maximum value. From **Fig.3-C**; the maximum value of S&M index on the curve can be founded when **F = 0.5**. Taking **F = 0.5** instead of the last empirical value of **F (0.95959595)** and repeating our proposed approach's manipulation steps on the same image shown in **Fig. 2**, we will obtain another 20 optimized results. The FOM's measurement values during the 20 results are plotted into **Fig.3-D**; where: the three colored lines; Blue, Green, and Red represents: in the three cases: Noisy, CVT, and WT; respectively.

E. Our Gray Phantom Study:

A new Gray phantom shown in **Fig. 4**; been introduced by *S. Badawy and M. GadAllah*; consists of gray scale vertical and horizontal bars used for testing strait boundaries and intensities of the gray levels. It is designed especially to test and assure the applied method for all gray scale images in radiology including Ultrasound's B-Scans.

Processing steps:

- a. Noising the phantom by an added noise determined by the **Snr** value.
- b. Denoising the produced image in Curvelet transform domain.
- c. Applying the wavelet based fusion method been explained previously in **B.**; in-between the noised and the denoised image produced from **a.** and **b.**, respectively taking **F = 0.5**.
- d. *Measurements:*
 - The edge detection map of Canny had been extracted as a qualitative measure of edge preservation for each image of the three images produced from **a.**, **b.**, and **c.**
 - A four quantitative performance evaluation measures PSNR, RMSE, Q, and FOM; had been calculated in-between the original phantom image and each one of its three derivatives produced in **a.**, **b.**, and **c.**
- e. Repeating the previous four steps **a.**, **b.**, **c.**, and **d.**; along **12** different values for **Snr**.

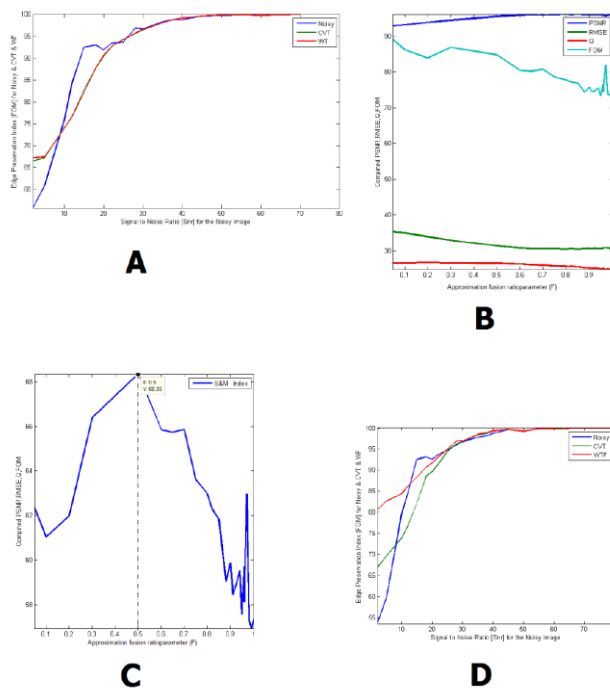


Fig. 3. Right Kidney Scan Image's Analytical Results Curves

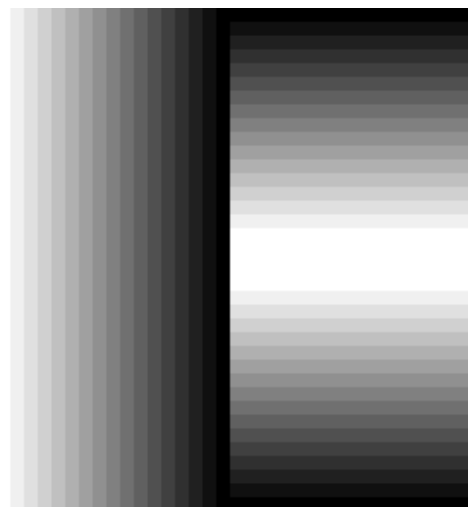


Fig. 4. Our New Gray Phantom

TABLE I. THE RIGHT KIDNEY SCAN AVERAGE RESULTS

| Average Results for: Final Wavelet Fused Image | Final Output Before S&M Index Optimization | Final Output After S&M Index Optimization |
|--|--|---|
| Peak Signal to Noise Ratio [PSNR] | 56.39 | 56.29 |
| Mean Square of Error [MSE] | 3.45 | 3.87 |
| Root Mean Square of Error [RMSE] | 1.26 | 1.31 |
| Universal Quality Index [Q] | 0.34 | 0.34 |
| Edge Preservation Index [Pratt's FOM] | 91.04 | 94.09 |

III. RESULTS

A. Kidney scan image Results with $F = 0.95959595$:

After applying our proposed approach on the right kidney scan's image, where; $F = 0.95959595$, and along 20 different value of Snr; the result is 20 by 3 images produced 60 resulted images beside the original one. After Canny edge map had been extracted producing another 60 result edge maps for each image beside the edge map of the original one.

The results was good but after being optimized; it became better; so, we satisfied here only by displaying the final average quantitative results of the 20 different results for: PSNR, RMSE, Q, and FOM, in comparison by the same average results after being optimized by the suitable fusion ratio based on applying S&M index enhancement, see **Table I**. Analytically; the FOM's 20 results had been plotted vs. the corresponding Snr, as shown in **Fig.3-A**.

B. Kidney scan image Results with $F = 0.5$:

After applying the S&M optimization on the right kidney scan's image, selecting $(F) = 0.5$, and along 20 different value of Snr; the result is 20 by 3 images produced 60 resulted images beside the original one. After Canny edge map had been extracted producing another 60 result edge maps for each image beside the edge map of the original one. A sample result from the 20 results, where $Snr = 10$; can be seen in **Fig. 5**; where: The lower four images in the two figures are the *canny edge map* for the upper four images correspondingly. Those maps are a qualitative measure for edge preservation in each image. The numbers on images are as follow:

- 1) Is stated for the Noisy-Image.
- 2) Is stated for the Reconstructed -Image after the denoising process in Curvelet transform domain.
- 3) Is stated for the Final Processed - Image after making the wavelet based image fusion in between 1 & 2.
- 4) Is stated for the Original Base Image as a reference.

The final average quantitative results of the 20 different results for: PSNR, RMSE, Q, and FOM, after being optimized by the suitable fusion ratio; are shown in Table I. The FOM's 20 results had been plotted vs. the corresponding Snr, as shown in **Fig.3-D**.

C. Our Phantom Study Results:

Applying our enhancement approach on the gray scale phantom image, selecting $(F) = 0.5$, and along 12 different value of Snr; the result is 12 by 3 images produced 36 resulted images beside the original one. After Canny edge map had been extracted producing another 36 result edge maps for each image beside the edge map of the original one. According to the tabulated results into **Table II**; we have up to twelve figures. A sample Result of our Gray Scale Phantom is shown in **Fig.6**; Where; $Snr = 25$ for the noisy image. Where: The lower three images are the *canny edge maps* for the upper three images correspondingly.

This is a qualitative measure for edge preservation in each image. The numbers on images are as follow:

- 1) Is stated for the Noisy-Image.
- 2) Is stated for the Reconstructed -Image after the denoising process in Curvelet Transform (CT) domain.
- 3) Is stated for the Final Processed - Image after making the wavelet based image fusion in between 1 & 2.

The quantitative measurements taken from processing our Gray Phantom image had been tabulated into Table II; representing 12 Results' measurements, taking $F_ratio = 0.5$. Table II has equal number of elements. Each row describes one complete result measurements for a specified Signal to Noise Ratio (SNR) value taken for the noised image. The last row is the mean or **Avg.** (Average) values of each parameter. The first-left column represent the value of SNR, the next column is the result number, the next 8 columns are the measurements values, Where;

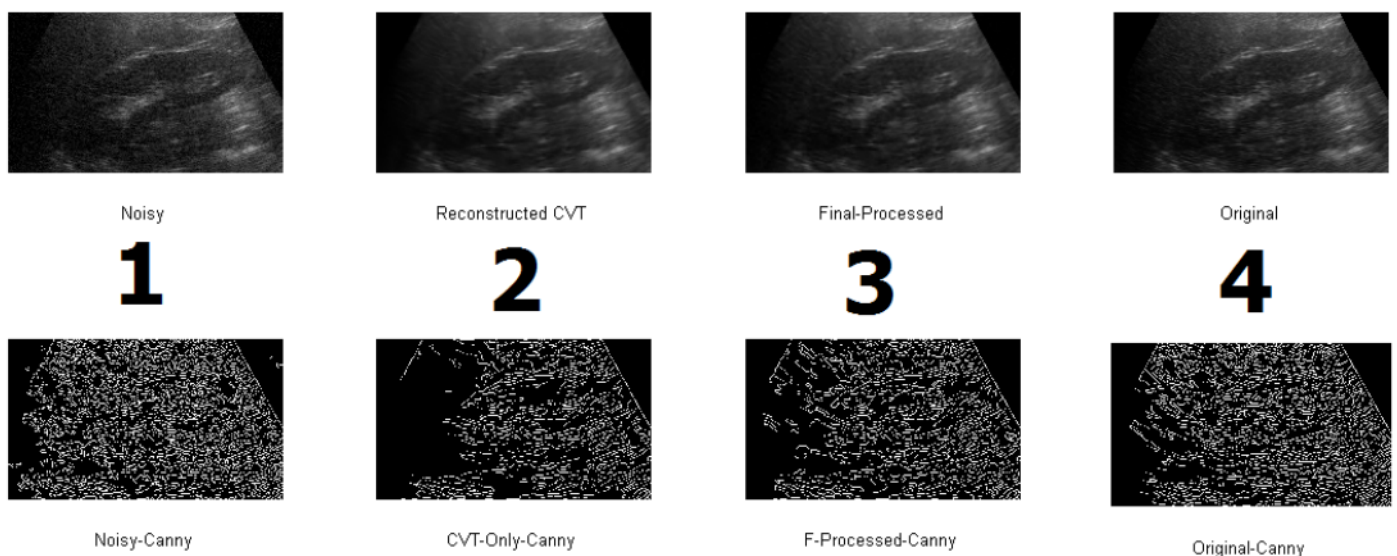


Fig. 5. A Sample Result for the Right Kidney Scan After Using S&M Index at Fusion Ratio, $F = 0.5$, Where; $Snr = 10$

Noisy - Snr: is the noisy image's *signal to noise ratio*.

Result - No.: is the *number of result*

Parameter + N: *N* refers to *Noisy* and means that this parameter is calculated in-between the noisy image and the reference original one shown in Fig.3.

Parameter + FP: *FP* refers to *Final Processed by the applied approach*, and means that this parameter is calculated in-between the *Final Result wavelet-fused image* and the reference original one.

The four calculated measures' 12 results; *PSNR, RMSE, Q, and FOM*; had been plotted for each image vs. the corresponding Snr; producing four curves shown in Fig.7: A, B, C, and D respectively. Where: The Blue line represents the quantity before applying the processing approach (Noisy) and the Green line represents the same quantity after image being processed (FP).

IV. DISCUSSION

Looking at Fig.3-A & Fig.3-D, obviously; we can see the enhancement in the edge preservation index FOM for the final fused kidney scan image, before and after applying S&M-optimization; represented by the red line.

Quantitatively, Looking at Table I, we can notice the increase in FOM after the optimization process, where the value of Q is still constant; assuring our optimized approach in preserving edges while maintaining quality as possible.

Qualitatively; comparing in between the Canny edge map for the final fused images shown in Fig.5-3, and the denoised image using Curvelet only shown in Fig.5-2; we can see the effect of our proposed approach in recovering image details which lost in denoising process. Comparing in between the Canny edge map for the final processed image shown in Fig.5-3, and the noisy image shown in Fig.5-1; we can see the enhanced denoising produced while keeping edges as possible.

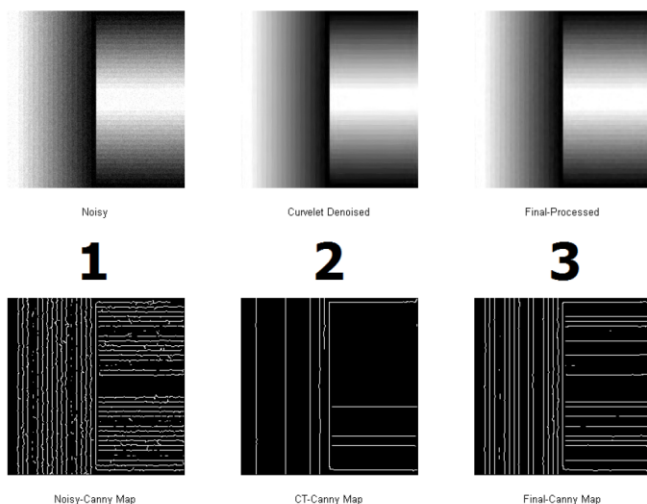


Fig. 6. A Sample Result of Our Gray Scale Phantom Where; Snr = 25

TABLE II. THE GRAY SCALE PHANTOM IMAGE'S RESULTS, WHERE; F= 0.5

| Noisy SNR | Result NO. | PSNR | | RMSE | | Q | | FOM | |
|----------------|------------|--------------|--------------|--------------|-------------|-------------|-------------|--------------|--------------|
| | | N | FP | N | FP | N | FP | N | FP |
| 2 | 1 | 6.07 | 22.53 | 126.81 | 19.06 | 0.01 | 0.19 | 35.34 | 49.29 |
| 5 | 2 | 9.07 | 25.29 | 89.78 | 13.86 | 0.02 | 0.28 | 35.35 | 51.14 |
| 8 | 3 | 12.07 | 27.76 | 63.56 | 10.44 | 0.03 | 0.38 | 36.95 | 57.22 |
| 10 | 4 | 14.07 | 29.36 | 50.49 | 8.68 | 0.05 | 0.45 | 37.83 | 60.13 |
| 12 | 5 | 16.07 | 30.71 | 40.10 | 7.44 | 0.07 | 0.52 | 39.07 | 67.97 |
| 15 | 6 | 19.07 | 32.44 | 28.39 | 6.09 | 0.11 | 0.60 | 43.91 | 76.58 |
| 18 | 7 | 22.07 | 34.04 | 20.10 | 5.06 | 0.19 | 0.67 | 55.46 | 69.92 |
| 20 | 8 | 24.07 | 35.15 | 15.97 | 4.46 | 0.26 | 0.72 | 61.81 | 80.68 |
| 25 | 9 | 29.07 | 38.23 | 8.98 | 3.13 | 0.50 | 0.81 | 87.24 | 78.58 |
| 30 | 10 | 34.07 | 43.20 | 5.05 | 1.76 | 0.72 | 0.89 | 93.51 | 75.97 |
| 40 | 11 | 44.07 | 49.62 | 1.60 | 0.84 | 0.89 | 0.91 | 93.97 | 97.11 |
| 50 | 12 | 54.07 | 57.61 | 0.50 | 0.34 | 0.91 | 0.91 | 100 | 100 |
| Average | | 23.65 | 35.50 | 37.61 | 6.76 | 0.31 | 0.61 | 60.04 | 72.05 |

Looking at Fig.7-A, Fig.7-B, and Fig.7-C; for PSNR, Q, and FOM, respectively; we can see analytically the resulted enhancement in the final processed gray scale image after applying our enhancement procedure; represented by the green line instead of the blue line before enhancement. Also, RMSE had been decreased through all the curve shown in Fig.7-D; represented by the green line instead of the blue line before.

Comparing in between the Canny edge map for the final fused phantom image shown in Fig.6-3, and the denoised image using Curvelet only shown in Fig.6-2; we can obviously; see the effect of our proposed approach in recovering some of image's details which lost in denoising process.

Comparing in between the Canny edge map for the final fused phantom image shown in Fig.6-3, and the noisy image shown in Fig.6-1; we can see qualitatively; the enhanced denoising produced by our proposed enhancement approach while keeping edges as possible preservation.

A clear improvement in the average value for FOM and Q can be shown quantitatively in Table II; to be 72.05 and 0.61 instead of 60.04 and 0.31, respectively.

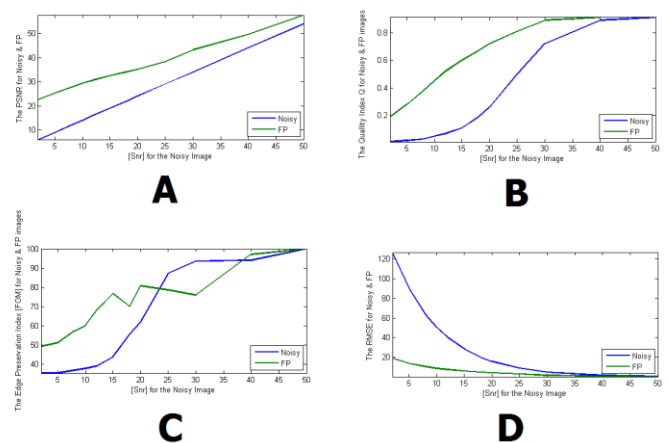


Fig. 7. Gray Scale Phantom Image's Analytical Results Curves

Also, the *average* value of PSNR was better enhanced after applying our proposed wavelet based approach to be **35.50** instead of **23.65** before.

On the other hand the *average* value of RMSE had been decreased to be **6.76** instead of **37.61** before; which assures that the applied approach is effective in *enhanced denoising with little values of error*.

The gray phantom results displayed is a step towards generalizing our introduced enhancement method for all gray scale diagnostic scans' images including Ultrasonography.

V. CONCLUSION

Wavelet based image fusion after Curvelet denoising, successfully used to *improve the removal of image's speckles (here, kidney's ultrasound scan) while enhancing its edges as possible*. Moreover, a gray phantom was introduced to help in testing and assuring the ability of the proposed work as an enhancement method for gray scale diagnostic scans' images including Ultrasonography. Also; a general quality optimization index be named S&M is newly introduced for selecting the best parameter for any parameter-based image fusion method being firstly introduced into image processing research.

ACKNOWLEDGMENT

Thanks to Dr. Mohammad Bahaa, *Consultant Radiologist*, Radiology Department, Specialized Clinic Center [SCC], Hawally, Kuwait; for his participation in the Practical work.

REFERENCES

- [1] Andrew G. Webb, "Ultrasonic Imaging", Ch. (3), in *Introduction to Biomedical Imaging*, © 2003 IEEE, Inc., IEEE Press Series on Biomedical Engineering, pp. 107-153, Published by: John Wiley & Sons, Inc., Hoboken, New Jersey ISBN: 0-471-23766-3
- [2] Shung, K. Kirk, "Gray-Scale Ultrasonic Imaging", Ch. 4, pp. 79-101, in *Diagnostic Ultrasound: Imaging and Blood Flow Measurements*, CRC Press, © 2006 by Taylor & Francis Group, LLC.
- [3] Torbørn Eltoft, "Modeling the Amplitude Statistics of Ultrasonic Images", IEEE T Med Imaging, Vol. 25, No. 2, Feb. 2006
- [4] K. Shahnazi and M. Fox, "Speckle Reduction in Real Time Ultrasound Imaging", © 1994 IEEE
- [5] Christoph B. Burckhardt, "Speckle in Ultrasound B-Mode Scans", IEEE Transactions on Sonics and Ultrasonics, Vol. SU-25, No. 1, Jan. 1978, © 1978 IEEE
- [6] Samir M. Badawy, Mohammed T. GadAllah and Mohammed M. Sharaf, "Intraocular Ultrasound Scan Examination by Image Segmentation", *Mitteilungen Klosterneuburg*, ISSN: 0007-5922, Volume: (64), Issue: (3), pp. 278-285, March 2014. URL: <http://mitt-klosterneuburg.com/show.php?v=64&i=3>
- [7] Mohammed T. GadAllah, M. M. Sharaf, Fahima A. Essawy and Samir M. Badawy, "Visual Improvement for Hepatic Abscess Sonogram by Segmentation after Curvelet Denoising", *International Journal of Image, Graphics and Signal Processing (IJIGSP)*, vol.5, no.7, pp.9-17, DOI: 10.5815/ijigsp.2013.07.02, June 2013. URL: <http://www.mecspress.org/ijigsp/ijigsp-v5-n7/IJIGSP-V5-N7-2.pdf>
- [8] Mohamed Tarek GadAllah & Samir Badawy, "Diagnosis of Fetal Heart Congenital Anomalies by Ultrasound Echocardiography Image Segmentation after Denoising in Curvelet Transform Domain", *Online Journal on Electronics and Electrical Engineering (OJEEE)*, ISSN (2090-0279), Vol. (5), No. (2), pp. 554 - 560, Ref. No. : W13-E-0023, April 2013. URL: <http://infomesr.org/attachments/W13-E-0023.pdf>
- [9] Mohamed T. GadAllah & Samir Badawy, "Aorta's abnormalities detection by Ultrasonography scan Denoising in Curvelet Transform Domain", Journal of Al-Azhar University Engineering Sector, JAUES, Vol. (7), No. (3), pp. 395 – 403, E38, Dec. 2012. *EJAUES website*: <https://sites.google.com/site/ejaues>
- [10] H. Lazrag, M. Ali H., and M. Saber N., "Despeckling of Intravascular Ultrasound Images using Curvelet Transform", *SETIT*, March 2012.
- [11] F. Y. Rizi, H. A. Noubari, and S. K. Setarehdan, "Wavelet-Based Ultrasound Image Denoising: Performance Analysis and Comparison", 33rd Conf. IEEE-EMBS, Boston, ©2011 IEEE
- [12] J. Starck, E. J. Candès, and D. L. Donoho, "The Curvelet Transform for Image Denoising", IEEE Transactions on Image Processing, Vol. 11, No. 6, June 2002, pp. 670-684, © 2002 IEEE
- [13] F. Yousefi Rizi, and S. K. Setarehdan, "Noise Reduction in Intravascular Ultrasound Images Using Curvelet Transform and Adaptive Complex Diffusion Filter: a Comparative Study", ICEE2012, Iran, 2012
- [14] M. Elhabiby ... et al, "Second Generation Curvelet Transforms Vs Wavelet transforms and Canny Edge Detector for Edge Detection from WorldView-2 data", *IJCSES*, Vol.3, No.4, Aug. 2012
- [15] E. J. Candès and D. L. Donoho, "Curvelets - A Surprisingly Effective Nonadaptive Representation for Objects with Edges", Saint-Malo Proceedings, Vanderbilt University Press, Nashville, TN, 1999, Available: <http://www-stat.stanford.edu/~donoho/Reports/1999/curveletsurprise.pdf>
Another URL: www.dtic.mil/cgi-bin/GetTRDoc?AD=ADP011978
- [16] David L. Donoho and Mark R. Duncan, "Digital Curvelet Transform: Strategy, Implementation and Experiments", Technical Report No. 2000-12, Department of Statistics, Stanford University, March, 2000. Available: <http://statweb.stanford.edu/~ckirby/techreports/GEN/2000/2000-12.pdf>
- [17] K. Ding, "Wavelets, Curvelets and Wave Atoms for Image Denoising", *3rd International CISP2010*, ©2010 IEEE
- [18] Unser, M., Aldroubi, A., and Laine, A., "Guest Editorial: Wavelets in Medical Imaging", *Special issue*, IEEE Transactions on Medical Imaging, Vol. 22, No. 3, pp. 285-288, March 2003
- [19] Y. Jin, E. Angelini, and A. Laine, "Wavelets in Medical Image Processing: Denoising, Segmentation, and Registration", Ch. 6, pp. 305-358, in *Handbook of Biomedical Image Analysis, Volume 1: Segmentation Models Part A*, By: J. S. Suri, D. L. Wilson, and S. L., © 2005 Kluwer Academic / Plenum Publishers, New York
- [20] R. C. Gonzalez, R. E. Woods, "Wavelets and Multiresolution Processing", Ch. 7 pp. 349-408, in *Digital Image Processing, 2nd Ed.*, © 2002 by Prentice-Hall, Inc., New Jersey 07458
- [21] Weaver, J. B. ... et al, "Filtering noise from images with wavelet transforms", *Magn. Reson. Med.*, Vol. 21, No. 2, pp. 288–295, 1991
- [22] A. Achim, A. Bezerianos, and P. Tsakalides, "Ultrasound Image Denoising Via Maximum A Posteriori Estimation of Wavelet Coefficients", *Proceedings of the 23rd Annual International Conference of the IEEE-EMBS*, Vol. 3, pp. 2553–2556, 2001
- [23] M. Misić, Y. Misić, G. Oppenheim, and J. Poggi, "Image Processing with Wavelets", Ch. 8, pp. 235-276, and, "An Overview of Applications", Ch. 9, pp. 279-310, in *Wavelets and their Applications*, © ISTE Ltd, London W1T 5DX, UK, 2007
- [24] James S. Walker, "Wavelet-Based Image Processing, *To Berlina*", *Applicable Analysis*, Vol. 85, No. 4, 2006, pp. 439-458, Available: <http://www.uwec.edu/walkerjrs/media/WBIP.pdf>
- [25] Øyvind Ryan, "Applications of the wavelet transform in image processing", Dep. of informatics, University of Oslo, 12 Nov. 2004
- [26] Adhemar Bultheel, "Wavelets with Applications in Signal and Image processing", Oct. 25, 2002, Available: <http://f3.tiera.ru/ShiZ/math/Signals/Bultheel/Wavelets%20with%20Applications.pdf>
- [27] R. Polikar, "The Story of Wavelets", © IMACS/IEEE CSCC'99 Proceedings, pp.5481-5486, 1999
- [28] M. Unser, and A. Aldroubi, "A Review of Wavelets in Biomedical Applications", Invited Paper, Proceedings of the IEEE, Vol. 84, No. 4, April 1996, pp. 626-638, © 1996 IEEE
- [29] A. Achim, A. Bezerianos, and P. Tsakalides, "Novel Bayesian Multiscale Method for Speckle Removal in Medical Ultrasound Images", *IEEE*

- Transactions on Medical Imaging*, Vol. 20, No. 8, pp. 772–783, Aug. 2001
- [30] A. Pizurica, W. Philips, I. Lemahieu, and M. Acheroy, "A Versatile Wavelet Domain Noise Filtration Technique for Medical Imaging", *IEEE Transactions on Medical Imaging*, Vol. 22, No. 3, pp. 323–331, March 2003
- [31] H. Rabbani ... et al, "Speckle Noise Reduction of Medical Ultrasound Images in Complex Wavelet Domain Using Mixture Priors", *IEEE-TBME*, Vol. 55, No. 9, Sep. 2008, ©2008 IEEE
- [32] Mohamed Ali Hamdi, "A Comparative Study in Wavelets, Curvelets and Contourlets as Denoising Biomedical Images", *IJ. Image, Graphics and Signal Processing*, 2012, 1, pp. 44-50, © 2012 MECS
- [33] Sandeep Palakkal, "Ridgelet and Curvelet first generation Toolbox", 25 May 2011, (Updated 21 Mar 2012), <http://www.mathworks.com/matlabcentral/fileexchange/31559-ridgelet-and-curvelet-first-generation-toolbox>
- [34] Anil A. Patil, and J. Singhai, "Image denoising using curvelet transform: an approach for edge preservation", *Journal of Scientific & Industrial Research*, Vol. 69, pp. 34-38, Jan. 2010
- [35] Tamilselvi, P.R., and P. T., "Noise suppression and improved edge texture analysis in kidney ultrasound images", *ICICT- 2010*
- [36] Ahmed Badawi, "Scatterer Density in Nonlinear Diffusion for Speckle Reduction in Ultrasound Imaging: The Isotropic Case", *World Academy of Science, Engineering and Tech.*, 13, 2008
- [37] Z. Yang, and M. D. Fox, "Speckle Reduction and Structure Enhancement by Multichannel Median Boosted Anisotropic Diffusion", *EURASIP Journal on Applied Signal Processing 2004:16*, pp. 2492–2502, © 2004 Hindawi Publishing Corporation
- [38] K. Z. Abd-Elmoniem, A. M. Youssef, and Y. M. Kadah, "Real-Time Speckle Reduction and Coherence Enhancement in Ultrasound Imaging via Nonlinear Anisotropic Diffusion", *IEEE-TBME*, Vol. 49, No. 9, pp. 997-1014, Sep. 2002, ©2002 IEEE
- [39] X. Zong ... et al, "Speckle Reduction and Contrast Enhancement of Echocardiograms via Multiscale Nonlinear Processing", *IEEE Transactions on Medical Imaging*, Vol. 17, No. 4, pp.532-540, 1998
- [40] Richard N. C., Douglas L. Jones, and W. D. O'Brien, "Ultrasound Speckle Reduction by Directional Median Filtering", © 1995 IEEE
- [41] R. Blum, Z. Xue, and Z. Zhang, "An overview of image fusion," Ch. 1, in *Multi-Sensor Image Fusion and Its Applications*, pp. 1–36, CRC Press Taylor and Francis Group, 2006
- [42] Paul M. de Zeeuw ... et al, "Multimodality and Multiresolution Image Fusion", *VISAPP 2012*, pp. 151-157
- [43] Paul M. de Zeeuw, "A Multigrid Approach to Image Processing", *R. Kimmel, N. Sochen, J. Weickert (Eds.): Scale-Space 2005*, LNCS 3459, pp. 396–407, 2005, © Springer-Verlag Berlin Heidelberg
- [44] Hui Li, B. S. Manjunath, and S. K. Mitra, "Multi-Sensor Image Fusion Using the Wavelet Transform", ©1994 IEEE
- [45] Zhou Wang and Alan C. Bovik, "Mean Squared Error: Love It or Leave It? A New Look at Signal Fidelity Measures", *IEEE Signal Processing Magazine*, Vol. 26, No. 1, pp. 98-117, Jan. 2009.
- [46] Zhou Wang, "Free Matlab based software program for calculating the Universal Image Quality Index; Q", Copyright (c) 2001 The University of Texas at Austin, Available: http://www.cns.nyu.edu/~zwang/files/research/quality_index/img_qi.m
The original source paper is: Zhou Wang, and Alan C. Bovik, "A Universal Image Quality Index", *Signal Processing Letters, IEEE*, ISSN: 1070-9908, Vol. 9, Issue: 3, pp. 81-84, DOI: 10.1109/97.995823, March 2002
- [47] W. K. Pratt, "Edge Detection", Ch. 15, pp. 443-508, in *Digital Image Processing: PIKS Inside, 3rd Ed.*, by William K. Pratt, ISBN: 0-471-22132-5, © 2001 by John Wiley & Sons, Inc.
- [48] Bjeny, "matlab source code for Edge preservation measure Index: Pratt's Figure of Merit", Programmers United Develop Net [PUDN], 23 Oct. 2009, Website: http://en.pudn.com/downloads201/sourcecode/graph/texture_mapping/detail947301_en.html
- [49] Ultrasound Console: ProSound 3500SX, Hitachi Aloka Medical, Ltd., URL: http://www.aloka.com/products/view_system.asp?id=3
- [50] Fusion of two images, image analysis, discrete wavelet analysis, Wavelet Toolbox, Documentation Center, MathWorks website, URL: <http://www.mathworks.com/help/wavelet/ref/wfusing.html>

AUTHORS PROFILE



Mohammed Tarek GadAllah was born in Menofia, Egypt on December-29th 1986. He is now working with a company; as a *Maintenance & Operating Engineer, Broadcast Division*, at; Filka Television Broadcasting Station - Filka Island – Kuwait. He is engineer on leave from *El Nasr Electric And Electronic Apparatus Co. [NEESAEE]*, Alexandria, Egypt. Mohammed was awarded the Degree of Bachelor of Electronic Engineering in July 29th, 2008 from *Industrial Electronics and Control Engineering Department*, Faculty of Electronic Engineering, Menofia University - Menofia, Egypt. Mohammed now has more than one achievement in designing and fabricating electronic circuits. Mohammed is now a *post-graduate – master student* at *Industrial Electronics and Control Engineering Department*, Faculty of Electronic Engineering, Menofia University, Egypt. Mohammed, as well as this paper, has four published papers: [6], [7], [8], and [9]. His previous and current research interests include: Biomedical Image Processing, Biomedical Engineering, Medical Imaging, Digital Control, and Electronic Circuits.



Samir Mohammed Badawy is now a *Doctor Emeritus, Department of Industrial Electronics and Control Engineering*, Faculty of Electronic Engineering, Menofia University - Menofia, Egypt. Samir had received his Ph.D. degree from Institute of cancer research, Royal Marsden Hospital, London University, UK. He had received his Master Degree from Helwan University, Egypt. He was awarded the Degree of Bachelor of Engineering from *Industrial Electronics Department*, Faculty of Engineering, Menofia University – Egypt. His previous and current researches interests include: Biomedical Physics, Biomedical Electronic, BCI, ECI, Medical Image Processing, Enhancement and Analysis, Ultrasound Tissue Characterization and Reconstruction, and biological Magnetic Effects on living cells.

A Second Correlation Method for Multivariate Exchange Rates Forecasting

Agus Sihabuddin, Subanar, Dedi Rosadi, Edi Winarko

Computer Science Graduate Program, Faculty of Mathematics and Natural Science,
Gadjah Mada University, Yogyakarta Indonesia

Abstract—Foreign exchange market is one of the most complex dynamic market with high volatility, non linear and irregularity. As the globalization spread to the world, exchange rates forecasting become more important and complicated. Many external factors influence its volatility. To forecast the exchange rates, those external variables can be used and usually chosen based on the correlation to the predicted variable. A new second correlation method to improve forecasting accuracy is proposed. The second correlation is used to choose the external variable with different time interval. The proposed method is tested using six major monthly exchange rates with *Nonlinear Autoregressive with eXogenous input (NARX)* compared with *Nonlinear Autoregressive (NAR)* for model benchmarking. We evaluated the forecasting accuracy of proposed method is increasing by 16.8% compared to univariate NAR model and slight better than linear correlation on average for D_{stat} parameter and gives almost no improvement for MSE.

Keywords—forecasting; foreign exchange; NARX; second correlation

I. INTRODUCTION

Exchange rate forecasting has proven to be predicted and univariate exchange rate forecasting gives a good forecast accuracy[1]. However, univariate specifications are limited. Those limitations are the market could be efficient and only driven from outside indicators, the available time series are too short for significant technical analysis with the chosen forecasting horizon[2], univariate model for some exchange rates do not provide a good forecast [3].

Multivariate time series analysis is an important statistical tool to study the behavior of time dependent data and forecast the future values depending on the historical data. With multivariate time series analysis, important dynamic inter-relationship among variables of interest like central bank interest rate, interest rate spread, and other exchange rates movements can be captured.

Research in exchange rates forecasting often uses daily, weekly or even monthly data. In multivariate exchange rates forecasting, monthly data is more often used than daily or weekly data due to the relationship with interest rate and inflation data that comes monthly [4], less volatile than the daily data and more referred to nonlinear data [5].

In this paper, we evaluate that the forecast accuracy could be improved by adding another external variable to capture that interrelationship by second correlation method. The advantage of this method is a simple technique to choose external variable

for a neural networks with still a good generalization capability [6]. A neural network algorithm is used because there is no need to specify a particular model and the model is adaptively formed based on the features presented from the data [1].

The rest of the paper is organized as follows. Section 2 discusses related work, Section 3 describes proposed correlation method, Section 4, presents experiment procedure and some experiment results, Section 5 presents the conclusion of the paper, and finally the future work is presented in Section 6.

II. RELATED WORK

Correlation is one of the most powerful tools for measuring linear relationship between two exchange rates or inter market relationships and evaluating inputs for neural networks [7]. Many research have elaborated correlation [2], [3], [8], [9] to measure relationships between series and evaluating input.

As the global economics keeps changing so the inter market connection changed too, so constant correlation to choose appropriate input may not be suitable. If input variables do not represents relevant ones, neural networks could not be expected to accurately predict the dependant variables or computational resources are wasted during training. In a case of a currency union like USDEUR there are divergent dynamics of real effective exchange rates in the individual country and form a new equilibrium [10].

III. SECOND CORRELATION

The research goals were sets as follows:

- 1) *Introduce to a second correlation to choose another variable to a neural networks algorithm.*
- 2) *Investigate the accuracy contribution of the method if it makes the accuracy increase or not.*

The second correlation described as follows:

- 1) *Compute the correlation coefficient for specific time interval if there is new exchange rates equilibrium, in this case the USDEUR came into the exchange rates market and change the equilibrium.*
- 2) *Use variable with highest correlation as input to Nonlinear Autoregressive with eXogenous input (NARX).*

In this case the second correlation is taken from January 2000 until 2014. Year 2000 is chosen as USDEUR currency entered the exchange rate market on January 1999; and Euro coins and Bank Note entered the market on January 2002. At that time, we assume there was a new equilibrium condition on

the exchange market by USDEUR. It is can be seen from the market share liquidity of USDEUR before 2001 was just near 0% and then since 2001 was 37.9% and became the second most liquid currency after USD (89.9%) [11]. The correlation with time interval of 2000 until 2014 of interest rate, interest rate spread and other exchange rates are used to choose external variable rather than using all time correlation data.

A NARX network is used to test the proposed method (M3) compared to usual correlation method (M2) for multivariate forecasting model, and *Nonlinear Autoregressive (NAR)* (M1) for univariate model.

IV. EXPERIMENTS AND RESULTS

A. Data

The exchange rates data used here are 40 years close price of major rates monthly data from January 1975 until April 2014. The dataset is long enough so we can see the effect of fundamental data on the market [7]. The major exchange rates used here are USDAUD, USDCAD, USDJPY, USDGBP, USDEUR and USDCHF which are the major exchange rates in the foreign exchange market and 65.2% of exchange market liquidity in April 2013 [11]. Each data contains 472 records which is divided into 80% (377 data) for training, 5% (24 data) for validation, and 15% (71 data) for testing, this data partition is similar to[12]–[15].

The external inputs or variables are the central bank interest rates, interest rate spread between two countries or region, and other exchange rates. The interest rates data are taken from Reserve Bank of Australian [16], Bank of Canada [17], De Nederlandsche Bank [18], Bank of England [19], Bank of Japan [20], Swiss National Bank [21], and Federal Reserve [22]. The interest rates and its spread correlations for six currencies do not give enough clues to be included in the external variables except for Japan interest rate and USDJPY (0.82). Correlation between interest rates and exchange rates is presented in Table I.

TABLE I. INTEREST RATE CORRELATION (1975-2014)

| | USDAUD | USDCAD | USDEUR | USDGBP | USDJPY | USDCHF |
|--------|--------------|-------------|-------------|--------------|-------------|--------------|
| AUD IR | -0.18 | | | | | |
| CAD IR | | 0.00 | | | | |
| EUR IR | | | 0.55 | | | |
| GBP IR | | | | -0.30 | | |
| JPY IR | | | | | 0.82 | |
| CHF IR | | | | | | -0.01 |
| US IR | | | | | | |

IR : Interest Rate

The interest rate spread correlation of two countries is presented in Table 2.

TABLE II. INTEREST RATE SPREAD CORRELATION (1975-2014)

| IRS | USDAUD | USDCAD | USDEUR | USDGBP | USDJPY | USDCHF |
|-------------|--------|--------|--------|--------|--------|--------|
| USD-AUD IRS | 0.54 | | | | | |
| USD-CAD IRS | | 0.25 | | | | |
| USD-EUR IRS | | | -0.15 | | | |
| USD-GBP IRS | | | | -0.11 | | |
| USD-JPY IRS | | | | | 0.54 | |
| USD-CHF IRS | | | | | | -0.01 |

IRS : Interest Rate Spread

Exchange rates correlation among six currencies gives two pairs of correlation those may be chosen for external variables. Those are USDAUD-USDCAD (0.83) and USDJPY-USDCHF (0.91). The exchange rates correlation among six countries or regions is presented in Table III.

TABLE III. EXCHANGE RATES CORRELATION (1975-2014)

| Correlation | USDAUD | USDCAD | USDEUR | USDGBP | USDJPY | USDCHF |
|-------------|--------|--------|--------|--------|--------|--------|
| USDAUD | 1.00 | 0.83 | 0.44 | 0.53 | -0.50 | -0.20 |
| USDCAD | | 1.00 | 0.57 | 0.46 | -0.19 | 0.07 |
| USDEUR | | | 1.00 | 0.68 | 0.31 | 0.57 |
| USDGBP | | | | 1.00 | -0.21 | 0.05 |
| USDJPY | | | | | 1.00 | 0.91 |
| USDCHF | | | | | | 1.00 |

B. Second Correlataion Process

The correlation coefficient between of interest rate and exchange rate from 2000 until 2014 (Table IV) does not give enough clues because of low value of correlation to make it as one of external variable.

TABLE IV. INTEREST RATE CORRELATION (2000-2014)

| IR | USDAUD | USDCAD | USDEUR | USDGBP | USDJPY | USDCHF |
|--------|--------|--------|--------|--------|--------|--------|
| AUD IR | 0.23 | 0.16 | 0.11 | -0.51 | 0.42 | 0.39 |
| CAD IR | 0.57 | 0.50 | 0.51 | -0.22 | 0.67 | 0.72 |
| EUR IR | 0.69 | 0.60 | 0.55 | -0.14 | 0.74 | 0.76 |
| GBP IR | 0.62 | 0.56 | 0.47 | -0.34 | 0.79 | 0.74 |
| JPY IR | -0.27 | -0.37 | -0.28 | -0.33 | -0.08 | -0.14 |
| CHF IR | 0.57 | 0.45 | 0.54 | -0.04 | 0.57 | 0.70 |
| US IR | 0.48 | 0.38 | 0.44 | -0.25 | 0.62 | 0.66 |

IR : Interest Rate

Correlation coefficient between interest rate spread and exchange rates from year 2000 until 2014 gives better correlation for USDJPY and USDCHF. The correlation data is presented in Table V.

TABLE V. INTEREST R ATE SPREAD CORRELATION (2000-2014)

| IRS | USDAUD | USDCAD | USDEUR | USDGBP | USDJPY | USDCHF |
|-------------|--------|--------|--------|--------|--------|--------|
| USD-AUD | 0.49 | | | | | |
| USD-CAD IRS | | 0.02 | | | | |
| USD-EUR IRS | | | -0.08 | | | |
| USD-GBP IRS | | | | -0.18 | | |
| USD-JPY IRS | | | | | 0.65 | |
| USD-CHF IRS | | | | | | 0.70 |

IRS : Interest Rate Spread

The correlation coefficient among exchange rates since 2000 until 2014 shows more currencies with stronger correlation. There are USDAUD-USDCAD, USDAUD-USDEUR, USDCAD-USDEUR, USDAUD-USDCHF, USDCAD-USDCHF, USDEUR-USDCHF, with correlation coefficient values above 0.9. These conditions support the assumption that the equilibrium of exchange rates market has changed to form more relationships among exchange rates. The correlation coefficient of exchange rates is presented in Table VI.

TABLE VI. EXCHANGE RATES CORRELATION (2000-2014)

| Correlation | USDAUD | USDCAD | USDEUR | USDGBP | USDJPY | USDCHF |
|-------------|--------|--------|--------|--------|--------|--------|
| USDAUD | 1.000 | 0.964 | 0.907 | 0.434 | 0.741 | 0.945 |
| USDCAD | | 1.000 | 0.903 | 0.429 | 0.695 | 0.911 |
| USDEUR | | | 1.000 | 0.600 | 0.569 | 0.913 |
| USDGBP | | | | 1.000 | -0.129 | 0.316 |
| USDJPY | | | | | 1.000 | 0.757 |
| USDCHF | | | | | | 1.000 |

The correlation coefficient of USDAUD-USDCAD at the periods of time 1975 to 2014 is 0.83. To forecast USDAUD, USDCAD should be used as external variable. In 2000-2014 periods, we have USDAUD-USDCAD, USDAUD-USDEUR, and USDAUD-USDCHF with high correlation values. The other USDAUD multivariate forecasting model may choose USDCHF as external variable instead of others. The plot data for USDAUD and USDCHF is presented in Fig. 1.

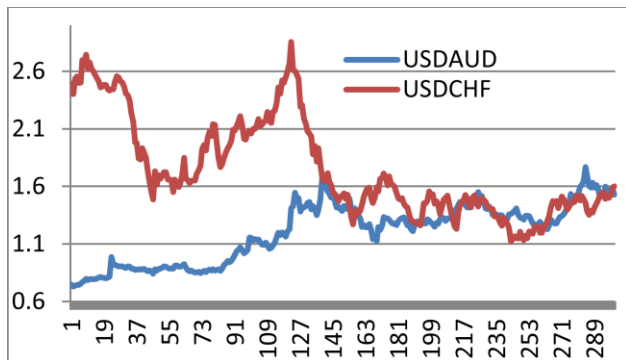


Fig. 1. Plot data for USDAUD and USDCHF (1975-1999)

The correlation coefficient of USDAUD-USDCHF is increasing from 2000 and can be seen in Fig. 2

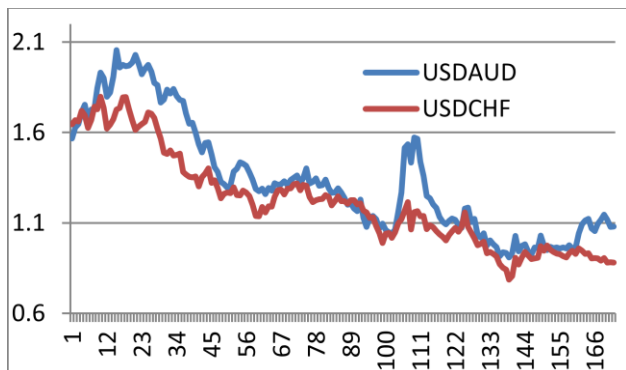


Fig. 2. Plot data for USDAUD and USDCHF (2000-2014)

C. Forecast Measure

In order to evaluate the forecast accuracy of the models, two forecast error measurements are used: Mean Square Error (MSE) and directional statistics (D_{stat}). MSE is defined as follows[23]:

$$MSE = \sum_{i=1}^n \frac{e_i^2}{n}$$

D_{stat} is defined as follows [3]:

$$D_{stat} = \frac{1}{N} \sum_{t=1}^N at * 100\%$$

where $at=1$ if $(x_{t+1} - x_t)(\hat{x}_{t+1} - x_t) \geq 0$ otherwise 0.

D_{stat} is more preferable in financial instruments forecasting because it gives the correctness of gradient prediction [3].

D. Result

The result of the proposed method presented in Table VII, VIII, and IX. NAR is used to get the univariate exchange rate forecasting and then compared to adding other variable with correlation from 1975 to 2014, and the other external variable added from the correlation from 2000 to 2014. The external variable for correlation from 1975 and correlation from 2000 for each currency is presented in Table VII.

TABLE VII. EXTERNAL VARIABLES FOR EXCHANGE RATES FORECASTING

| External | USDAUD | USDCAD | USDEUR | USDGBP | USDJPY | USDCHF |
|----------|--------|--------|--------|--------|--------|--------|
| M2 | USDCAD | USDAUD | USDGBP | USDEUR | USDCHF | USDJPY |
| M3 | USDCHF | USDCHF | USDCHF | USDAUD | USDAUD | USDAUD |

M2 :Model with 1975-2014 correlation

M3 :Model with 2000-2014 correlation

The result comparison for MSE of proposed method is presented in Table VIII. Our result shows that adding a new variable on model M2 and M3 from six major exchange rates does not gives MSE improvement significantly except for M3 in USDCHF (-84.87%).

TABLE VIII. MSE PERFORMANCE OF UNIVARIATE AND MULTIVARIATE

| Ext. Var | USDAUD | USDCAD | USDEUR | USDGBP | USDJPY | USDCHF |
|----------|-----------------------|----------------------|----------------------|----------------------|--------------------|-----------------------|
| M1 | 0.003587 | 0.001138 | 0.000667 | 0.000326 | 7.5655 | 0.001375 |
| M2 | 0.003302 (-7.93%) | 0.001101 (-3.16%) | 0.000659 (-1.06%) | 0.000325 (-0.44) | 8.1223 (7.36%) | 0.001411 (2.60%) |
| M3 | 0.003155 (-12.05%) | 0.001082 (-4.93) | 0.00066 (-1.08) | 0.000336 (-3.17%) | 7.3374 (-3.02%) | 0.000208 (-84.87%) |

The results for D_{stat} accuracy parameter is more promising in M2 model with best significant accuracy improvement for USDJPY (23.53%); and in M3 model give more significant accuracy improvement for USDCHF (41.46%). For both MSE and D_{stat} accuracy parameter, USDEUR does not provide good forecasting improvement.

TABLE IX. D_{STAT} PERFORMANCE OF UNIVARIATE AND MULTIVARIATE

| Ext. Var. | USDAUD | USDCAD | USDEUR | USDGBP | USDJPY | USDCHF |
|-----------|--------------------|--------------------|-------------------|--------------------|--------------------|--------------------|
| M1 | 57.75% | 47.89% | 60.56% | 52.11% | 47.89% | 57.75% |
| M2 | 61.97% (7.31%) | 52.11% (8.82%) | 61.97% (2.33%) | 61.97% (18.92%) | 59.15% (23.53%) | 60.56% (4.88%) |
| M3 | 64.79% (12.19%) | 56.34% (17.65%) | 61.97% (2.33%) | 60.56% (16.22%) | 60.53% (26.47%) | 81.69% (41.46%) |

In overall result, in MSE accuracy parameter does not give significant accuracy improvement. In D_{stat}, it can be seen that this proposed method gives better accuracy with average of 64.08%, higher than univariate model 53.99% accuracy and usual correlation method with 59.62% accuracy. It gives a 19.39% accuracy improvement on average with the highest accuracy is achieved at USDCHF with 81.69% accuracy.

Other advantage of using the second correlation method is it gives a simple way to choose an external variable rather than using other more complex algorithm or by trial and error to choose one by one other series.

V. CONCLUSION

The second correlation method does not give MSE improvement significantly. From D_{stat} it can be seen that this method gives better accuracy with average of 64.08%, higher than univariate model 53.99% accuracy and usual correlation method with 59.62% accuracy. It gives a 19.39% accuracy improvement on average with the highest accuracy is achieved at USDCHF with 81.69% accuracy. It does not give better accuracy for USDEUR in MSE and D_{stat} .

VI. FUTURE WORK

Our future work will continue to implement the method in the different exchange rates; and combine the method with other such technical indicator.

REFERENCES

- [1] G. P. Zhang, "Time Series Forecasting using A Hybrid ARIMA and Neural Network Model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.
- [2] B. R. Setyawati, R. C. Creese, and M. Jaraiedi, "Neural Networks for Univariate and Multivariate Time Series Forecasting," *Proceeding 2003 IIE Annu. Conf.*, 2003.
- [3] J. Yao and C. L. Tan, "A Case Study on Using Neural Networks to Perform Technical Forecasting of Forex," *Neurocomputing*, vol. 34, no. 1–4, pp. 79–98, Sep. 2000.
- [4] B. Majhi, M. Rout, R. Majhi, G. Panda, and P. J. Fleming, "New Robust Forecasting Models for Exchange Rates Prediction," *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12658–12670, 2012.
- [5] M. C. Medeiros and A. Veiga, "A Hybrid Linear-Neural Model for Time Series Forecasting," *IEEE Trans. Neural Networks*, vol. 11, no. 6, pp. 1402–12, Jan. 2000.
- [6] S. Samarasinghe, *Neural Networks for Applied Sciences and Engineering*. New York: Auerbach Publication, 2007.
- [7] B. R. Setyawati, "Multi-Layer Feed Forward Neural Networks for Foreign Exchange Time Series Forecasting," *Dissertation*, West Virginia University, 2005.
- [8] W. Nan and H. Tieshan, "Study on Correlation between Different NDF Data and Fluctuations of RMB Exchange Rate," *Int. J. Econ. Financ.*, vol. 5, no. 5, pp. 55–63, 2013.
- [9] F. H. Nordin, F. H. Nagi, and A. A. Z. Abidin, "Comparison Study of Computational Parameter Values Between LRN and NARX in Identifying Nonlinear Systems," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 21, pp. 1151–1165, 2013.
- [10] A. Rusek, "The Eurozone 's Equilibrium Real Exchange Rates," *Mod. Econ.*, vol. 3, no. September, pp. 534–541, 2012.
- [11] Bank for International Settlements, "Triennial Central Bank Survey Foreign Exchange Turnover in April 2013 : Preliminary Global Results," 2013.
- [12] K. Kim and W. B. Lee, "Stock Market Prediction Using Artificial Neural Networks with Optimal Feature Transformation," *Neural Comput. Appl.*, vol. 13, pp. 255–260, 2004.
- [13] I. Kaastra and M. Boyd, "Designing A Neural Network for Forecasting Financial and Economic Time Series," *Neurocomputing*, vol. 10, pp. 215–236, 1996.
- [14] M. Al Mamun and K. Nagasaka, "Artificial Neural Networks Applied to Long-term Electricity Demand Forecasting," in *Proceedings of the Fourth International Conference on Hybrid Intelligent Systems (HIS'04)*, 2004, pp. 0–5.
- [15] A. M. Oyewale, "Evaluation of Artificial Neural Networks in Foreign Exchange Forecasting," *Am. J. Theor. Appl. Stat.*, vol. 2, no. 4, pp. 94–101, 2013.
- [16] Reserve Bank of Australia, "Monthly Interest Rate 2014," 2014. [Online]. Available: www.rba.gov.au. [Accessed: 06-Apr-2014].
- [17] Bank of Canada, "Monthly Interest Rate," 2014. [Online]. Available: www.bankofcanada.ca. [Accessed: 06-Apr-2014].
- [18] De Nederlandsche Bank, "Monthly Interest rate," 2014. [Online]. Available: www.dnb.nl. [Accessed: 06-Apr-2014].
- [19] Bank of England, "Monthly Interest Rate," 2014. [Online]. Available: www.bankofengland.co.uk. [Accessed: 06-Apr-2014].
- [20] Bank of Japan, "Monthly Interest Rate," 2014. [Online]. Available: www.boj.or.jp. [Accessed: 06-Apr-2014].
- [21] Swiss National Bank, "Monthly Interest Rate," 2014. [Online]. Available: www.snb.ch. [Accessed: 06-Apr-2014].
- [22] Federal Reserve, "Monthly Interest Rate," 2014. [Online]. Available: www.federalreserve.gov. [Accessed: 06-Apr-2014].
- [23] J. G. De Gooijer and R. J. Hyndman, "25 Years of Time Series Forecasting," *Int. J. Forecast.*, vol. 22, no. 3, pp. 443–473, Jan. 2006.

Mitigation of Cascading Failures with Link Weight Control

Hoang Anh Tran Quang
Dept. of Computer Science
National Defense Academy of Japan
Yokosuka, Kanagawa, Japan

Akira Namatame
Dept. of Computer Science
National Defense Academy of Japan
Yokosuka, Kanagawa, Japan

Abstract—Cascading failures are crucial issues for the study of survivability and resilience of our infrastructures and have attracted much interest in complex networks research. In this paper, we study the overload-based cascading failure model and propose a *soft defense strategy* to mitigate the damage from such cascading failures. In particular, we assign adjustable weights to individual links of a network and control the weight parameter. The information flow and the routing patterns in a network are then controlled based on the assigned weights. The main idea of this work is to control the traffics on the network and we verify the effectiveness of the load redistribution for mitigating cascading failure. Numerical results imply that network robustness can be enhanced significantly using the relevant smart routing strategy, in which loads in the network are properly redistributed.

Keywords—cascading failures; link's weight; network robustness

I. INTRODUCTION

Nowadays, many complex systems in nature and society can be described by intricate network patterns, including technological, social and biological systems such as the Internet, the World-Wide Web, electrical power grid networks, metabolic networks and so on. In recent years, complex network research has also attracted much attention and becomes an useful tool for scientists to make major advances in understanding salient properties of complex human engineered systems, that go beyond the single component behaviour. A vast number of studies have clarified that certain topological properties of complex networks have strong impacts on their stability. An early important work of Albert, Jeong and Barabasi [1] showed that scale-free networks which have heterogeneous degree distributions, are remarkably resistant against random errors, but at the same time, targeted malicious attacks can easily disrupt the networks by removing only a small fraction of nodes or links. On the other hand, homogeneous degree distribution networks – namely, random networks, might be considerably stable against attacks but somewhat vulnerable to random failures.

Since a vulnerability is a weakness which might reduce a system performance, recently, one of the major focuses of complex network research, is the vulnerability management. In our daily life, cascading failures are common phenomenon and can occur in many natural and man-made systems, due to endogenous or exogenous (or can be both in some cases) factors.

There are many types of cascading failures that are mentioned, from some critical infrastructures such as electrical power grids and computer networks, to economic, ecological, even political systems. A common yet hard-to-predict property of cascading failures is that even a single point of failure emerges locally, the damage is widely propagated and could result in global collapse.

In decades, a number of important aspects of cascading failures in complex networks have been discussed and many valuable results have been found. There several works studied the impact of cascading failures on different types of power grid networks such as the North American power grid network [2], the European power grid network [3], and the Italy power grid network [4]. Other works studied cascading failures in other types of complex networks, such as telecommunication networks [5], or socio-technological networks [6]. As we further model and understand the behaviour of cascading failures, how to build in safeguards that may be able to prevent them in the future, has become a central topic of interest.

Available set of existing methods to enhance network robustness against cascading failures can be generally divided into two classes

- A set of methods to improve network robustness *statically*, which has been developed in order to prevent cascading failures before the occurrence of initial failures.
- A set of methods to improve network robustness *dynamically*, which has been developed in order to minimize the damage of cascading failures after some initial failures occurred.

An example study of the former is the paper of Shin and Namatame [7]. In their paper, they considered network robustness and design cost as a trade-off function and used an evolutionary algorithm to evolve networks. Their results revealed that clustering, modularity, and long path lengths all play an important part in the design of robust large-scale infrastructure.

Typical examples of the latter include the well-known method proposed by Motter [8]. In his paper, he introduced and investigated a costless defense method based on a selective removal of nodes and edges immediately after initial failure and showed that the proposed method is practical and can drastically reduce the size of the cascade.

The main idea in [8] is that a selective set of *insignificant* nodes that process little but contribute relatively large loads to the network are removed so as to reduce the overall load in the network. This approach has the advantage of a low incremental investment cost, as it requires the ability to perform a remote shutdowns of nodes. However, it also has a strong disadvantage since it is difficult to provide early detection of cascading failures and it requires knowledge of the global topology.

There are essentially two types of strategies for defending or mitigating cascading failures

- *Hard strategy* to prevent cascading failures. This type of strategy has a disadvantage of impacting the topology of networks.
- *Soft strategy* to minimize the damage of cascading failures without any change in the connection of networks.

Both of the above-mentioned methods [7, 8] can be regarded as *hard strategy* type. While the latter shows its disadvantage in directly impacting to the topological structure of networks, the former may become a harder strategy since the purpose is to design robust networks from the beginning while it has been showed that most of networks in reality already have their own specific existing structures.

To overcome the difficulties of *hard strategy*, some *soft strategies* to counter cascading failures without impacting to the connections of given networks, have been recommended. Wang and Kim [9], Li, Wang, Sun, Gao, and Zhou [10] developed new capacity models to cascading failures to make the network more robust, while at the same time the cost to assign capacities is drastically reduced. Meanwhile, in the survey of Chen, Huang, Cattani, and Altieri [11], they reviewed strategies for improving transport efficiency, including soft strategies to design efficient routing strategies and also hard strategies to adjust the underlying network structure.

Because *hard strategies* are not always applicable in many cases, we mainly focus on *soft control strategy* in this work. Among existing literature, the most related work to ours is the paper by Yang, Wang, Lai, and Chen [12]. In their paper, they discovered an optimal solution to both cascading failures and traffic congestion problem. They provided numerical evidence and theoretical analysis to show that, by choosing a proper weighting parameter, a maximum level of robustness against cascades and traffic congestion can be simultaneously achieved. However, the critical tolerance parameters which are the minimal values to prevent cascading failures that they showed in their paper are applied for all nodes in the network. It implies that, to prevent overload in some nodes, they unexpectedly increased the capacities in other nodes which are may be unnecessary and become waste redundancy, and of course it leads to much cost. Besides, they did not consider the connectivity of the network after initial failures, which is a relevant index in studying network robustness.

In this paper, we control load distribution in a network via several smart routing schemes. We define network robustness in considering the connectivity of the network.

We evaluate network robustness to capture the effectiveness of the proposed method on an artificial generated scale-free network and some realistic networks subjected to intentional attacks. Simulation results show the significant enhancement of network robustness when a smart routing strategy is adapted.

The reminder of this work is organized as follows: we first present the cascading failure model in Section II. We then introduce the proposed routing strategy and simulation settings in Section III and IV. We present numerical results in Section V, and finally summarize this work in Section VI.

II. A CASCADING FAILURE MODEL

Cascading breakdown in complex networks is regarded as an avalanching failure, where the failure of a few local nodes can result in a global-scale breakdown of the network. In various types of existing cascading failures, one of the most prominent cascade phenomenon that occurs in most infrastructure networks, is overloaded cascading failure.

This type of cascading failures can take place in electrical power grid networks, when for any reason a line breaks down, its power is automatically shifted to the neighbouring lines. In most of the cases, the neighbouring lines can handle the extra load, but sometimes, these lines are also overloaded and continuously shift their load to their neighbours. This eventually leads to a cascade of failures where a large number of transmission lines are overloaded and malfunction at the same time. For instance, due to the power redistribution, some typical incidents have taken place in history, such as the blackout on August 14, 2003 when an initial disturbance in Ohio led to the largest blackout in the history of the United States and millions of people throughout parts of North Eastern and Mid-Western United States, and Ontario, Canada, were without power for as long as 15 hours [13].

Furthermore, the overloaded cascading failures can also take place on the Internet, where traffic is rerouted to bypass breakdown routers, eventually leading to an avalanche of overloads on other routers which do not have enough capability to handle extra traffic, and a large drop in the performance. A prominent example is the congestion on the early Internet in October 1986, when the NSFnet phase-I backbone dropped three orders of magnitude from its capacity of 32 kbit/s to 40 bit/s, and this continued to occur until end nodes started implementing Van Jacobson's congestion control between 1987 and 1988 [14].

The interesting feature of this type of overloaded cascading failures is that it does not necessarily propagate through adjacent physical contact, i.e. the single failure of one node in a network may cause failures to non-adjacent nodes due to the network's load redistribution. The potential impact of this type of cascading breakdown on the security of large complex networks, has been firstly investigated by Motter and Lai [15].

Since traffic or information is usually transmitted along the shortest paths in most communication networks, it has been suggested that the information flow across the network – namely the load L , can be captured well by the betweenness centrality, which can be calculated as the number of shortest paths that pass through a node when flow is sent from each

available generation node to each distribution node (load in unweighted networks)

$$L = \text{shortest path betweenness.} \quad (1)$$

We consider the networked system with N nodes. The possibility of observing cascading failures is enabled by assigning flow capacities to each of the nodes of the system. Here, the capacity of a node is defined as the maximum load that the node can handle. Since engineered systems are optimized for maximum capacity and minimum cost, it is assumed that the capacity of the nodes is proportional to the initial load [15, 16]

$$C_i = \alpha L_i(0), \quad i = 1, 2, \dots, N \quad (2)$$

where C_i is the capacity of node i , $L_i(0)$ is the initial load of node i which is defined in (1). The tolerance parameter α ($\alpha \geq 1$) captures the relationship between network component capacity and load demand levels. Here, the tolerance parameter α also implies the budget of network construction or resource allocation.

Suppose that $s_i(t)$ represents the state of node i at time step t . A very simple condition to recognize that node i will fail or not at time step t is the following relation

$$s_i(t) = \begin{cases} 1, & \text{if } L_i(t) > C_i \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $s_i(t) = 1$ indicates that node i will fail at time step t , and $s_i(t) = 0$ indicates that node i will be safe.

Each disruptive event triggers flow redistribution within the networks and can potentially lead to cascading failures. Initially, a network is in a stationary state in which the load at each node is smaller than its capacity. It is possible that from some reasons a breakdown occurs at one or more nodes – some nodes in the system is overloaded beyond the given capacity, so that they cannot work at all, and can be assumed that be removed from the network, causing the change of transmission paths in the network. The breakdown of one or some heavily loaded nodes will cause the redistribution of loads over the remaining nodes, which can trigger breakdowns of newly overloaded nodes. These additional failures require a new redistribution of loads, which either stabilizes and the failures are locally absorbed, or grows until a large number of nodes are compromised to a failure point.

Using the model, we are able to follow the dynamical response of the system to failures, and in particular to model how the failure in one location can propagate and have consequences over the whole network. The model is applicable to many realistic situations in which the flow of physical quantities in the network, as characterized by the loads on nodes, is important.

III. SMART ROUTING STRATEGY

Any network can be represented by an adjacency matrix A .

The element of matrix A in the i^{th} row and the j^{th} column is expressed as a_{ij} . If $a_{ij} = 1$, node i and node j are connected, and if $a_{ij} = 0$, these two nodes are not connected.

We assume that a weight of an arbitrary link connecting a node i and j is assigned proportionally to the connectedness of the two nodes as follows

$$w_{ij} = a_{ij}(k_i k_j)^\beta \quad (4)$$

where k_i and k_j is the degree – the number of links, of node i and j , respectively, and β is the weight control parameter.

The introduced weight of a link connecting a node i and j can be also referred to as resistance of the link against the flow, which is determined, for example, by the conductance in an electrical network. As it can be observed in (4), the control parameter $\beta > 0$ indicates that links connecting hubs – node with high degrees, have high weights, and will be avoided using to transmit information. This assumption matches the reality since lines that have high resistances will obstruct the flow in networks. In contrast, $\beta < 0$ implies that low weights are assigned to links connecting hubs, meaning that these links have low resistances and are frequently used to transmit information. The final regime $\beta = 0$ corresponds to the case in which all links have the equal weight (same resistance). In this case, the flow will be transmitted by a predetermined rule, e.g. via the shortest paths in networks. It is worth noting that the weights we assign for links in networks here, are only dummy values – these values do not correspond to any measurement in reality, e.g. the geographical distance between two cities, the resistance of a transmission line between two substations, etc. The idea of this work is to control the flow of communication in networks based on these dummy values.

The weight of a path from a node m to node n , that passing through a set of l intermediate nodes $S = \{1, 2, \dots, l\}$ is the total link weights including in the path

$$w_{m \rightarrow n} = \sum_{i=1}^{l-1} w_{ij}, \quad j = i + 1 \quad (5)$$

from which, the shortest path on the weighted network, within all possible weighted paths between m and n can be obtained.

As introduced in section II, the shortest path based betweenness of a node i can be used as an approximation of the load that flows through i . Nevertheless, this definition of load has the disadvantage that is only applicable for unweighted networks. Based on the mentioned weight in (4), we extend the definition of load that is also applicable for weighted networks.

In particular, the load of a node i can be approximated by the total number of shortest *weighted* paths that pass through that node (load in weighted networks)

$$L = \text{shortest weighted path betweenness.} \quad (6)$$

IV. SIMULATION SETTINGS AND PERFORMANCE MEASUREMENT

We conduct simulations with both artificial generated and realistic networks. We use a scale-free network generated by Barabasi-Albert model [17] as a benchmark network, which has the number of nodes $N = 1000$ and the average degree $\langle k \rangle = 4$. We use realistic networks such as: the Euro-road network – a road network located mostly in Europe where nodes represent cities and a link denotes that nodes are connected by a road [18, 19]; the Western States of the United States of America – a node is either a generator, a transformer or a substation, and a link represents a power supply line [18, 20]; the autonomous system peering information inferred from Oregon route-views between March 31, 2001 and May 26, 2001 [21]; the network of e-mail interchanges between members of the University Rovira I Virgili [18, 22]; and the top 500 busiest commercial airports in the United States [23, 24]. These networks have been chosen in order to represent a wide variety of complex network topologies. Additional statistical information of the networks used in this paper is summarized in Table 1, where N is the number of nodes; E is the number of links; $\langle k \rangle$ is the average degree; and k_{max} is the maximum degree.

We first show the effect of the weight control parameter β in (4) to the load distribution of the benchmark scale-free network in Fig. 1.

As shown in the figure, by adjusting the weight control parameter β , the scale-free network discloses its load distributions in different manners while its topological structure is kept unchanged. $\beta = 0$ corresponds to the case where all links in the network are assigned an equal weight ($w_{ij} = 1$, for all i, j). In this case, the scale-free network shows its heterogeneous load distribution – the higher degree a node has, the higher load it carries, since all nodes tend to use hubs as shortcuts to transmit information along the network. If we reduce the parameter β to -1 , we obtain the most heterogeneous load distribution among three cases. In this case, low weights are assigned to links connecting hubs, meaning that hubs are more and more frequently used to transmit information. As expected, $\beta = 1$ shows the most homogeneous load distribution where links connecting hubs will be avoided using to transmit information. In this case, hubs experience a significant decrease in load. On the other hand, nodes which carried a small load, may acquire a larger one. In other words, all nodes contribute equivalently to transmitting information along the network.

If a node has a relatively small load, its removal will not cause major changes in the load balance, and subsequent overload failures are unlikely to occur. However, when the load at a node is relatively large, its removal is likely to significantly affect loads at other nodes and possibly start a sequence of overload failures. To study the attack vulnerability of a network, the procedure for selecting the order in which nodes are removed is an open choice. A tractable choice, used in the original study of complex networks, is based on aiming at the most connected nodes, and highest loaded nodes. This is a deterministic process since the topology of the network is known at every point in time. To explore the effects of our

proposed method, only nodes disrupted at intentional attacks are included in the analysis. The node with the largest load L_{max} is chosen for node attacks, and L_{max} is recalculated after every node removal when more than one element is eliminated according to the intensity of the disruptive events.

TABLE I. STATISTICAL INFORMATION OF NETWORKS USED IN THIS PAPER

| Network | Category | N | E | $\langle k \rangle$ | k_{max} |
|---------------|----------------------|-------|-------|---------------------|-----------|
| Scale-free | Artificial generated | 1000 | 1997 | 3.99 | 72 |
| Euro-road | Physical | 1174 | 1417 | 2.41 | 10 |
| US power grid | Physical | 4941 | 6594 | 2.67 | 19 |
| Internet | Physical | 10670 | 22003 | 4.12 | 2312 |
| E-mail | Communication | 1133 | 5451 | 9.62 | 71 |
| Top 500 | Physical | 500 | 2980 | 11.92 | 145 |

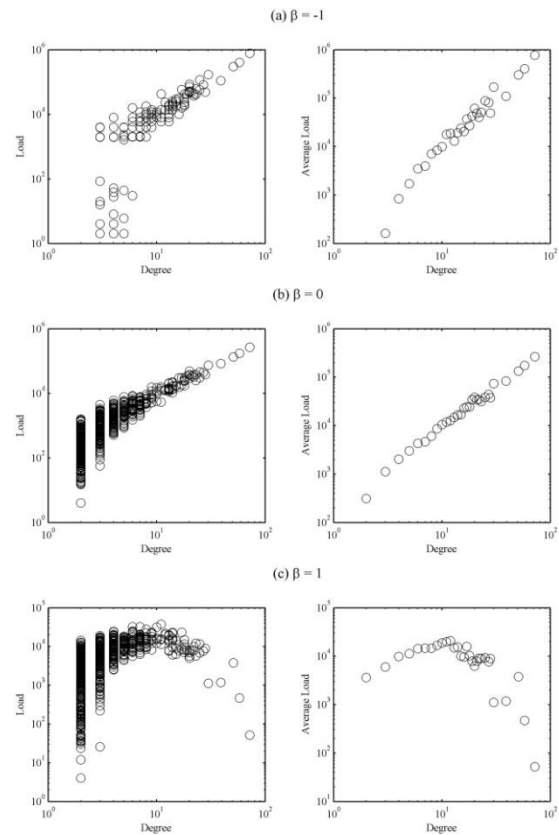


Fig. 1. The relation of load distribution vs. degree and average load vs. degree of scale-free network with the weight control parameter (a) $\beta = -1$, (b) $\beta = 0$, (c) $\beta = 1$.

The evaluation of robustness focuses on some generic topological metrics of network such as the size, efficiency, and average shortest path length of the *Largest Connected Component LCC* – the component for which there is a path

between any pair of nodes in a network. In addition to considering properties of the *LCC*, some other metrics are also considered, e.g., the average avalanche size and distribution, the critical point of phase transition from an absorbing to cascading state. Since the connectivity of the system is important, it is reasonable to consider the *LCC* as network robustness.

In this paper, we quantify the network robustness using R , the ratio of functional nodes in the *LCC* before and after the cascading event caused by the failure of a single node with highest load

$$R = N' / N \quad (7)$$

where N and N' are the sizes of the *LCC* of the network before and after cascading failure event, respectively. Evidently, N is the size of the initial network and $0 \leq R \leq 1$. A network has high integrity if $R \approx 1$, i.e., there is no cascade in the network and all nodes are almost fully connected and functional after initial failure. Otherwise, $R \approx 0$ indicates that a network has been disintegrated into several small sub-networks. Thus, the relative size of R is appropriate for representing the robustness of a network to cascading failures. Using the model presented in Section II and this definition of network robustness, we obtain the familiar property "robust yet fragile" for which, in scale-free networks, R remains close to unity in the case of random breakdowns, but is significantly reduced under attacks that target nodes with the highest loads.

V. SIMULATION RESULTS

Intuitively, the most effective and simple method to prevent a cascading failures is to increase the tolerance parameter α so that all nodes have sufficient resources to prevent failure due to overload. Another solution is redistributing load of a failure node.

The resulting networks provide information about the minimum capacity that each remaining node must be able to carry to survive without triggering cascade. The capacity that a node i must have for preventing cascade at any initial one node failure, is the maximum overall networks with single removal: $C_i = \max_{j \in N_i} L_i(N_i \setminus j)$, where $L_i(N_i \setminus j)$ is the load on the node i in the network with the node j removed. However, the capacity is often limited by cost thus it is impractical to assign sufficient large capacity to all nodes in networks. Based on this fact, and also to validate the effectiveness of our method, we assume that the tolerance parameter α is taken as $1 \leq \alpha \leq 2$, implying that there is no much redundant capacity in the system. We evaluate the efficiency of our proposed approach for small value of α , and show we can mitigate cascading failures without needing to increase the capacity of each node.

Since the difficulty of early detection makes the reactive defense strategy after initial attack but prior to the cascade an unrealistic damage control strategy for many real-world networks, we focus on the scenario of seeking an appropriate routing strategy before initial failures, indicating the static proactive defense strategy where we design a robust routing strategy against predicted attacks a priori.

Fig. 2 shows the network robustness defined in (7) with the assumption of only a single node with the highest load is failed initially.

The Euro-road and US power grid network are more likely random network, i.e. the degree and load distribution of the networks are homogeneous. On the other hand, the scale-free, Internet, E-mail and Top 500 have the degree and load heterogeneously distributed. It is obviously shown in Fig. 2 that network robustness can be enhanced for all values of weight control parameter β if we simply increase the tolerance parameter α to allocate as much capacities as possible to nodes. However, it also exhibits that without considering a proper parameter β , the enhancement is not noticeable even when we have sufficient large α – a little change in the value of the weight control parameter may leads to the dramatic decrease of network performance. It implies that, to enhance network robustness significantly, we have to consider adjusting properly both tolerance parameter α and weight control parameter β . Simulation results show that the relation between the weight control parameter β and the tolerance parameter α strongly impacts to network robustness, and this relation is irregular for each individual network. In particular, as shown in the figure, we are able to archive high network robustness for

- Scale-free network with: $\beta \geq 0.5, \alpha \geq 1.3$.
- Top 500 airports network with: $0.5 \leq \beta \leq 0.7, \alpha \geq 1.5$.
- E-mail network with: $0 \leq \beta \leq 0.6, \alpha \geq 1.3$.
- The Internet with: $0 \leq \beta \leq 1, \alpha \geq 1.6$.
- Euro-road network with: $-0.75 \leq \beta \leq -0.5, \alpha \geq 1.5$.
- US power grid network with: $0.2 \leq \beta \leq 0.5, \alpha \geq 1.7$.

Fig. 2 shows the similar tendency of overwhelming hot color area where $\beta > 0$ compared with other area ($\beta \leq 0$) for heterogeneous networks. It indicates that we can significantly enhance network robustness against intentional attacks by choosing a proper routing strategy with $\beta > 0$, which transforms a network from heterogeneous state to homogeneous one.

Interestingly, we obtain different results for Euro-road and US power grid although they are both homogeneous networks. While network robustness may be enhanced due to some positive values of β in the case of US power grid network, the result in Euro-road network shows a different manner, i.e. network robustness is enhanced significantly with some negative values of β .

An evident truth emerges when β is small for all networks. With these β , load distribution of a network becomes extreme heterogeneous, and a single attack to a single highest load node may disrupt the whole system.

One more interesting result is also observed with some large values of β , in which network robustness start to decrease. We can explain this tendency as follows: with some intermediate values of β , a network is transformed from heterogeneous load distribution to a more homogeneous one gradually.

However, too large β leads a homogenized network once again becomes heterogeneous load distributed – nodes with small degrees become very high load nodes, and this makes intentional attacks devastate the system.

As shown in Fig. 2, we can classify strategies that enhance network robustness into four following classes

- Hub avoidance strategy ($\beta > 0$): efficient for scale-free network, top 500 airports network, the Internet.
- Hub oriented strategy ($\beta < 0$): efficient for Euro-road network.
- Strategy that increases the tolerance parameter: efficient for E-mail network.
- Strategy of both hub avoidance ($\beta > 0$) and increase of the tolerance parameter: efficient for US power grid network.

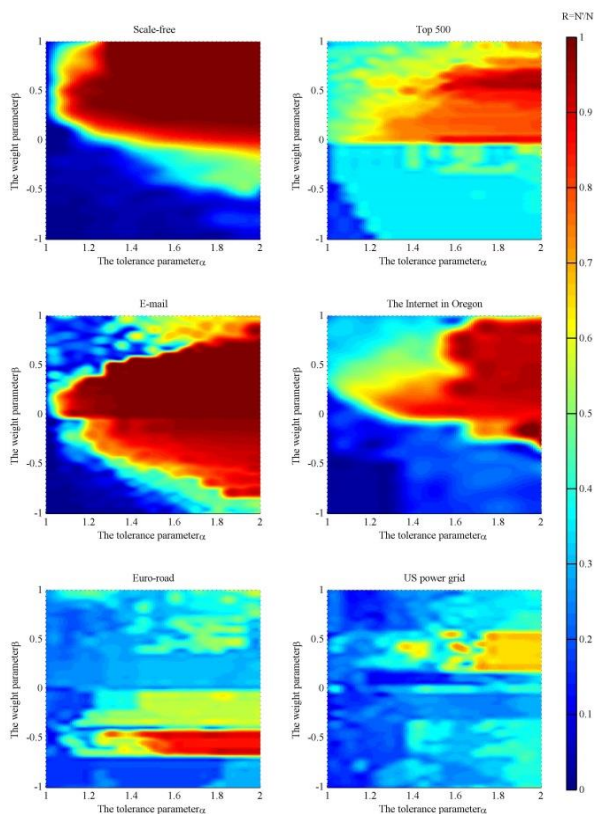


Fig. 2. Network robustness of the scale-free, Euro-road, US power grid, E-mail and Internet as the function of the tolerance parameter α in (2) and weight control parameter β in (4). In the figure, hot colors show the area of high robustness and cold colors correspond to the rest. In this scenario, we intentionally seek for an efficient design of routing strategy as a proactive defense strategy.

VI. CONCLUSIONS

In this paper, we proposed a routing strategy to mitigate the damage of cascading failures caused by overload. We assigned weights to links in networks and control the weight by an

adjustable parameter. Numerical results showed the effectiveness and the availability of our proactive method for critical infrastructure networks such as electrical power grid networks, the Internet, and so forth. Routing traffic in this manner can limit the damage of cascades by turning a heterogeneous load distribution into a more homogeneous one, reduces the need to shutdown nodes to stop a cascade, and simultaneously lowers the investment costs in network capacity layout.

For simplicity, in this paper, we assigned a weight to a link connecting two nodes, a value that proportional to the connectedness of the two nodes. However, almost systems in reality have more complicated, even unpredictable links weights. In addition, there are several alternative possibilities to the node load for the case in which the physical quantity of interest (information, packets, electric power, etc) does not travel through shortest paths only. Therefore, our future work is to investigate the two questions: how to logically assign weights to links of a network; and how to determine general flow manner. Thereto, infrastructure systems are becoming more interdependent and failures within a given system are more likely to reduce the performance of other systems [25, 26, 27, 28, 29]. Hence, how to mitigate cascading failures in such interdependent networks becomes an indispensable issue and will be also our future work.

REFERENCES

- [1] R. Albert, H. Jeong, A. Barabási, “Error and Attack Tolerance of Complex Networks”, *Nature*, 406, pp. 378–382, 2000.
- [2] R. Albert, I. Albert, G. L. Nakarado, “Structural Vulnerability of the North American Power Grid”, *Phys. Rev. E*, 69(2), 2005.
- [3] M. Rosas-Casals, S. Valverde, R. V. Sole, “Topological Vulnerability of the European Power Grid under Errors and Attacks”, *International Journal of Bifurcation and Chaos*, 17(7), pp. 2465–2475, 2007.
- [4] P. Crucitti, V. Latora, M. Marchiori, “A Topological Analysis of the Italian Electric Power Grid”, *Phys. A: Statistical Mechanics and its Applications*, 338(1), pp. 92–97, 2004.
- [5] M. F. Habib, M. Tornatore, F. Dikbiyik, B. Mukherjee, “Disaster Survivability in Optical Communication Networks”, *Computer Communications*, 36(6), pp. 630–644, 2013.
- [6] C. Barret, K. Channakeshava, F. Huang, J. Kim, A. Marathe, M. V. Marathe, G. Pei, S. Saha, S. P. Subbiah, A. K. S. Vullikanti, “Human Initiated Cascading Failures in Societal Infrastructures”, *PLoS ONE*, 7(10), 2012.
- [7] S. Y. Shin, A. Namatame, “Evolutionary Optimized Networks and Their Properties”, *International Journal of Computer Science and Network Security*, Vol. 9, No. 2, pp. 4–12, 2009.
- [8] A. E. Motter, “Cascade Control and Defense in Complex Networks”, *Phys. Rev. Lett.*, Vol. 93, 2004.
- [9] B. Wang, B. J. Kim, “A High Robustness and Low Cost Model for Cascading Failures”, *EPL* Vol. 78, No. 4, 2007.
- [10] P. Li, B. H. Wang, H. Sun, P. Gao, T. Zhou, “A Limited Resource Model of Fault-Tolerant Capability against Cascading Failure of Complex Network”, *The European Physical Journal B* 62(1), pp 101–104, 2008.
- [11] S. Chen, W. Huang, C. Cattani, G. Altieri, “Traffic Dynamics on Complex Networks: A Survey”, *Mathematical Problems in Engineering*, 732698, 2012.
- [12] R. Yang, W. X. Wang, Y. C. Lai, G. Chen, “Optimal Weighting Scheme for Suppressing Cascades and Traffic Congestion in Complex Networks”, *Phys. Rev. E* 79, 026112, 2009.
- [13] J. Glanz, R. Perez-Pena, “90 Seconds that Left Tens of Millions of People in the Dark”, *New York Times*, 2003.

- [14] V. Jacobson, "Congestion Avoidance and Control", in ACM SIGCOMM '88, Stanford, CA, pp. 314–329, 1988.
- [15] A. E. Motter, Y. C. Lai, "Cascade-based Attacks on Complex Networks", Phys. Rev. E.66, 2002.
- [16] P. Crucitti, V. Latora, M. Marchiori, "Model for Cascading Failures in Complex Networks", Phys. Rev. E 69, 045104, 2004.
- [17] A. Barabási, R. Albert, H. Jeong, "Scale-free Characteristics of Random Networks: the Topology of the World-Wide Web," Phys. A. Vol. 281, pp. 69–77, 2000.
- [18] Network dataset – KONECT, <http://konect.uni-koblenz.de/networks>, 2014.
- [19] L. Subelj, M. Bajec, "Robust Network Community Detection using Balanced Propagation", European Phys. Jour. B.81, pp. 353–362, 2011.
- [20] D. J. Watts, S. H. Strogatz, "Collective Dynamics of Small-world Networks", Nature. 393, No. 6684, pp. 440–442, 1998.
- [21] J. Leskovec, J. Kleinberg, C. Faloutsos, "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations", International Conference on Knowledge Discovery and Data Mining, 2005.
- [22] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, A. Arenas, "Self-similar Community Structure in a Network of Human Interactions", Phys. Rev. E.68, 2003.
- [23] Tore Opsahl Network Dataset, <http://toreopsahl.com/datasets>.
- [24] V. Colizza, R. Pastor-Satorras, A. Vespignani, "Reaction-Diffusion Processes and Metapopulation Models in Heterogeneous Networks", Nature Phys. 3, pp. 276–282, 2007.
- [25] S. E. Chang, H. A. Seligson, R. T. Eguchi, "Estimation of the Economic Impact of Multiple Life-line Disruption: Memphis Light, Gas, and Water Division Case Study", Technical Report No. NCEER-96-0011. Multidisciplinary Center for Earthquake Engineering Research, Buffalo, New York, 1996.
- [26] P. Pederson, D. Dudenhoeffer, S. Hartley, M. Permann, "Critical Infrastructure Interdependency Modeling: a Survey of US and International Research", Report INL/EXT-06-11464, Idaho Falls: Idaho National Laboratory, 2006.
- [27] Y. Y. Haimes, B. M. Horowitz, J. H. Lambert, J. R. Santos, C. Lian, K. G. Crowther, "Inoperability Input-Output Model for Interdependent Infrastructure Sectors I: Theory and Methodology", Journal of Infrastructure Systems, 11(2), pp. 67–79, 2005.
- [28] L. Dueñas-Osorio, J. I. Craig, B. J. Goodno, "Seismic Response of Critical Interdependent Networks", Earthquake Engineering and Structural Dynamics, pp. 285–306, 2007.
- [29] C. D. Brummitt, R. M. D. Souza, E. A. Leicht, "Suppressing Cascades of Load in Interdependent Networks", Proc. of the National Academy of Sciences, 2012.

Applicability of the Maturity Model for IT Service Outsourcing in Higher Education Institutions

Victoriano Valencia García
Computer Management Technician
and Researcher at Alcalá University
Madrid, Spain

Dr. Eugenio J. Fernández Vicente
Professor at Computer Science Dept.
Alcalá University
Madrid, Spain

Dr. Luis Usero Aragonés
Professor at Computer Science Dept.
Alcalá University
Madrid, Spain

Abstract—Outsourcing is a strategic option which complements IT services provided internally in organizations. This study proposes the applicability of a new holistic maturity model based on standards ISO/IEC 20000 and ISO/IEC 38500, and the frameworks and best practices of ITIL and COBIT, with a specific focus on IT outsourcing.

This model allows independent validation and practical application in the field of higher education. In addition, this study allows to achieve an effective transition to a model of good governance and management of outsourced IT services which, aligned with the core business of universities, affect the effectiveness and efficiency of its management, optimizes its value and minimizes risks.

Keywords—IT governance; IT management; Outsourcing; IT services; Maturity model; Maturity measurement

I. INTRODUCTION

One thing to change about ICT at university level is the deeply rooted approach which exists, or which used to exist, called infrastructure management. This kind of management has evolved into a governance and management model more in line with the times, which is a professional management of services offered to the university community [6]. It is for this reason that in recent years a set of methodologies, best practices and standards, such as ITIL, ISO 20000, ISO 38500 and COBIT, have been developed to facilitate ICT governance and management in a more effective and efficient way.

These methodologies, which are appropriate and necessary to move from infrastructure management to service management, see a lack of academic research. For that reason it is inadvisable to use these frameworks on their own, and it is advisable to consider other existing frameworks in order to extract the best from each for university level [6].

ICT or IT services have implications for business and innovation processes and may be a determinant in their evolution. The organization of these services, their status within the organization of the university, and their relationships with other management departments and new technologies is therefore vital. At present, the degree of involvement, the volume of services offered, and the participation or external alliances with partner companies through outsourcing, that Gottschalk and Solli-Saether [7] defined as the “practice of turning over all or part of an organization’s IT function to an IT vendor”, are of special interest.

Currently, and in the years to come, organizations that achieve success are and will be those who recognize the benefits of information technology and make use of it to boost their core businesses in an effective strategic alignment, where delivery of value, technology, risk management, resource management, and performance measurement of resources are the pillars of success.

It is necessary to apply the above-mentioned practices through a framework and process to present the activities in a manageable and logical structure. Good practice should be more strongly focused on control and less on execution. They should help optimize IT investments and ensure optimal service delivery. IT best practices have become significant due to a number of factors, according to COBIT [10]:

- Business managers and boards demanding a better return from IT investments;
- Concern over the generally increasing level of IT expenditure;
- The need to meet regulatory requirements for IT controls in areas such as privacy and financial reporting, and in specific sectors such as finance, pharmaceutical and healthcare;
- The selection of service providers and the management of service outsourcing and acquisition;
- Increasingly complex IT-related risks, such as network security;
- IT governance initiatives that include the adoption of control frameworks and good practices to help monitor and improve critical IT activities to increase business value and reduce business risk;
- The need to optimise costs by following, where possible, standardised, rather than specially developed, approaches;
- The growing maturity and consequent acceptance of well-regarded frameworks, such as COBIT, IT Infrastructure Library (ITIL), ISO 27000 series on information security-related standards, ISO 9001:2000 Quality Management Systems—Requirements Capability Maturity Model ® Integration (CMMI), Projects in Controlled Environments 2 (PRINCE2) and

A Guide to the Project Management Body of Knowledge (PMBOK); and

- The need for organizations to assess how they are performing against generally accepted standards and their peers (benchmarking)

Clearly ICTs have become ubiquitous in almost all organizations, institutions and companies, regardless of the sector to which they belong. Hence, effective and efficient ICT management to facilitate optimal results is necessarily essential.

Furthermore, in this environment of total ICT dependency in organizations using ICTs for the management, development and communication of intangible assets, such as information and knowledge [16], organizations become successful if these assets are reliable, accurate, safe and delivered to the right person at the right time and place, according to ITGI [10]. Also, knowledge integration mechanisms are important in helping knowledge utilization in client firms [18].

In short, Fernández [6] proposes that the proper administration of ICT will add value to the organization, regardless of its sector (whether social, economic or academic) and will assist it in achieving its objectives and minimizing risk.

Given the importance of proper management of ICT, the search for solutions to the alignment of ICT with the core business of organizations has accelerated in recent years. The use of suitable metrics or indicators for measurement and valuation, generate confidence in the management teams. This will ensure that investment in ICTs generates the corresponding business value with minimal risk [6].

The above solutions are models of good practice, metrics, standards and methodologies that enable organizations to properly manage ICTs. And public universities are not outside these organizations, though they are not ahead. In addition, interest in adopting models of governance and management of appropriate ICTs is not as high as it should be.

Two of the factors through which IT best practices have become important is, the selection of appropriate service providers and the management of outsourcing and procurement of IT services.

IT outsourcing has brought potential benefits in addition to many examples of the great organizational losses associated with this practice. Even with awareness of the potential for failure, the IT outsourcing industry continues to grow, as organizations communicate their desire to engage in IT outsourcing and their determination to decipher a method that enables successful IT outsourcing relationships [17].

In addition, a maturity model is a method for judging whether the processes used, and the way they are used, are characteristic of a mature organization [4].

Models by phases or levels allow us to understand how IT management strategies based on computing evolve over time [11]. According to these models, organizations progress through a number of identifiable stages. Each stage or phase

reflects a particular level of maturity in terms of IT use and management in the organization.

There are many maturity models in the literature, and they are applied to various fields, such as project management, data management, help desk, systems safety engineering. Most of them refer to either Nolan's original model [14] or the Capability Maturity Model of Software Engineering Institute (Carnegie Mellon Software Engineering Institute). The latter model describes the principles and practices underlying software processes, and is intended to help software organizations evolve from ad-hoc, chaotic processes to mature, disciplined software processes.

Nolan was the first to design a descriptive stage theory for planning, organizing and controlling activities associated with managing the computational resources of organizations. His research was motivated by the theoretical need for the management and use of computers in organizations. From 1973 until today, technology and the way it is used has changed a lot, but Nolan's original idea is still valid, and it will remain as long as the quality of services provided internally in organizations, or by external suppliers, is essential.

II. LITERATURE REVIEW ON MATURITY MODELS FOR IT OUTSOURCING AND COMPARISON CHART

Very few models or frameworks of IT outsourcing can be found in the literature, either from the point of view of the client or outsourcer. The few models or frameworks that exist are varied. After a thorough literature review, and taking into account the point of view of the customer, the following models have been found to be relevant:

- [M1] Managing Complex IT Outsourcing – Partnerships (2002) [2];
- [M2] Information Technology Outsourcing (ITO) Governance: An Examination of the Outsourcing Management Maturity Model (2004) [4];
- [M3] A Unified Framework for Outsourcing Governance (2007) [5];
- [M4] IT Outsourcing Maturity Model (2004) [1];
- [M5] Outsourcing Management Framework Based on ITIL v3 Framework (2011) [12];
- [M6] Multisourcing Maturity Model (2011) [6];
- [M7] Maturity model for IT outsourcing relationships (2006) [8];
- [M8] IT Governance Maturity and IT Outsourcing Degree: An Exploratory Study (2007) [3]; and
- [M9] Global Multisourcing Strategy: Integrating Learning From Manufacturing Into IT Service Outsourcing (2011) [13]

The following table shows the maturity models and frameworks above, along with the key areas or determinants that they are based on. All key areas shown in Table I form the basis of the maturity model designed for IT service outsourcing.

TABLE I. EXISTING MATURITY MODELS AND FRAMEWORKS ON IT
OUTSOURCING

| Key areas or determinants | Maturity models and frameworks about IT outsourcing | | | | | | | | |
|--|---|----|----|----|----|----|----|----|----|
| | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 |
| Formal Agreement | X | | | X | | X | X | | X |
| Service Measurement | | X | X | | X | X | | X | |
| Quality Management | | X | | | | | X | | |
| Monitoring and Adjustments | | X | X | | X | X | | X | |
| Alignment IT-Business | X | X | X | | | | | X | |
| IT Governance Structure | X | | X | | | | | X | |
| Service Level Agreement (SLA) | X | X | | X | X | | X | | |
| IT Service Registration | | | | | | | | | |
| Incident and Problem Management | | | | X | X | | | | |
| Changes | | | | | X | | | | |
| Testing and Deployment | | | | | | | | | |
| Control of External Providers | X | X | | | X | X | X | X | X |
| Business Risk | | X | X | | X | | | X | |
| Financial Management | | | | | X | X | X | X | X |
| Legislation | | | X | | | X | | | |
| Demand and Capacity Management | | | | | | | | | |
| Formal Agreement Management | X | | | | | | | | X |
| Knowledge Management | X | | X | | | | X | | |
| Guidelines on outsourcing an IT service (life cycle) | | | | | | | | | |

Taking into account all key areas shown in Table I, a holistic maturity model (henceforth MM) has been designed with a specific focus on IT outsourcing governance and IT service management. The model establishes where organizations involved in the study are in relation to the following control criteria and information requirements according to Cobit: effectiveness; efficiency; confidentiality; integrity; availability; compliance; and reliability. Other criteria, from the perspective of managing critical IT resources, are the following: applications; information; infrastructure; and people.

With regard to IT governance, standard ISO/IEC 38500:2008, published in 2008, aims to provide a framework of principles for directors of different organizations in order to manage, evaluate, and monitor the efficient, effective and acceptable use of information and communication technologies. The direction, according to ISO / IEC 38500, must govern IT in three main areas:

- **Management.** Direct the preparation and implementation of strategic plans and policies, assigning responsibilities. Ensure smooth transition of projects to production, considering the impacts on the operation, the business and infrastructure. Foster a culture of good governance of IT in the organization.
- **Evaluation.** Examine and judge the current and future use of IT, including strategies, proposals and supply agreements (both internal and external).
- **Monitoring.** Monitor IT performance measuring systems in order to ensure that they fit as planned.

According to the results of the "IT Governance Study 2007" [19][20], reasons compelling governments to create an IT structure in the university include: aligning IT objectives with strategic objectives; promoting institutional vision of IT; ensuring transparency in decision-making; cost reduction; increased efficiency; and regulation and compliance audits.

On service management, MM takes into account ISO/IEC 20000 and ITIL v3, but it is customized to integrate governance and management into a single model. The model moves towards an integration that facilitates the joint use of frameworks efficiently. Thus, the MM designed consists of five levels, with each level having a number of general and specific characteristics that define it. These are determined by the selection of general concepts that underpin the MM (see first column in Table I). The selection is always justified and countersigned by ISO 20000 and ISO 38500 standards and ITIL and COBIT best practice methodologies.

III. MATURITY MODEL PROPOSED

In order to design the proposed maturity model, we studied in detail every reference on the provision of IT services that there is in the ISO 20000 and ISO 38500 standards and ITIL v3 and COBIT methodologies. In addition, we investigated the relevant literature and failed to find any maturity model that brings together the previous methodologies with a specific focus on IT outsourcing. As a result, a number of concepts and subconcepts were categorized to form the basis of the maturity model.

The MM follows a stage structure and has two major components: maturity level and concept. Each maturity level is determined by a number of concepts common to all levels.

Each concept is defined by a number of features that specify the key practices which, when performed, can help organizations meet the objectives of a particular maturity level. These characteristics become indicators, which, when measured, determine the maturity level.

The MM defines five maturity levels: initial or improvised; repeatable or intuitive; defined; managed and measurable; and optimized.

The model proposes that organizations under study should ascend from one level of maturity to the next without skipping any intermediate level. In practice, organizations can accomplish specific practices in upper levels. However, this does not mean they can skip levels, since optimum results are unlikely if practices in lower levels go unfulfilled.

IV. METRICS FOR MATURITY MEASUREMENT

We have designed an assessment tool along with the maturity model that allows independent validation and practical application of the model. Therefore, the maturity of an organization indicates how successfully all practices that characterize a certain maturity level have gone fulfilled. The questions used in the questionnaire form the basis of the assessment instrument. They were extracted from each of the indicators defining each of the general concepts and key areas of the maturity model.

These general concepts and defining characteristics have been extracted from the following standards and methodologies:

- Standard ISO/IEC 20000 and methodology of good practices ITIL v3. Both provide a systematic approach to the provision and management of quality IT services.
- Standard ISO/IEC 38500:2008 provides guiding principles for directors of different organizations to manage, evaluate, and monitor the use of information and communication technologies effectively and efficiently.
- Cobit business-oriented methodology provides good practice through a series of domains and processes, as well as metrics and maturity models in order to measure the achievement of the objectives pursued.

In addition, new indicators have been developed based on the proposed model in order to assess appropriate aspects not reflected either in previous methodologies and standards or in the existing literature (e.g. the inclusion of service performance in the SLA and the use of user-satisfaction surveys in IT-business alignment).

To evaluate the maturity model of an organization using the model and the measurement instruments proposed, it is necessary to obtain a series of data resulting from the responses to the questionnaire based on the indicators that define the general concepts of our maturity model.

Table II shows one of the nineteen key areas or concepts that are the basis of the MM. The first column of the table shows the level or levels corresponding to the indicator located in the second column. The second column shows the survey questions and indicators for each of the questions or part of the questions. Finally, the third column shows the source where the indicator or item has been extracted as a feature of the general concept or key area of the model.

TABLE II. METRICS TABLE AND QUESTIONNAIRE

| Level | Code – Indicator – Question of Questionnaire | Source |
|---|---|--------------------|
| <i>Concept: Service Level Agreement (SLA)</i> | | ISO 20000 & ITIL |
| 2 | SLA1 - SLA - There is an SLA for each outsourced IT service provided by the service provider | ISO 20000 & ITIL |
| | SLA2 - Elements of SLA - SLAs include: | |
| 2 | SLA2a - Service availability | |
| 5 | SLA2b - Service performance | Myself |
| 3 | SLA2c - Penalties for breach of SLA | |
| 2 | SLA2d - Responsibilities of the parties | |
| 3 | SLA2e - Recovery Times | |
| 4 | SLA2f - Quality Levels | |
| 4 | SLA2g - Security requirements | |
| 3 | SLA3 - Frequency reviewing of SLA - SLAs are reviewed periodically at predefined intervals | ISO 20000 & myself |

Therefore, the maturity level of every higher education institution studied is measured by evaluating its development in each key area or concept, which is indicated by responses to items or indicators in metrics tables (see Table II). In order to qualify for a specific maturity level, the university surveyed must carry out all key practices of that level successfully.

V. OBJECTIVES OF THE MATURITY MODEL

The main purpose of the model is to fulfill as many requirements of an ideal maturity model for IT outsourcing in the governance and management of outsourced IT services as possible. With the identification and definition of some key concepts and an assessment tool, the model allows a systematic and structured assessment of organizations. Although the assessment instrument has a lot of qualitative responses, it also has quantitative responses, such as the degree of compliance with certain characteristics that define the maturity model (e.g. the degree of influence of the KPIs and KGIs in the penalties for breach of agreements).

The identification of key areas and concepts specifying its characteristics to constitute the underlying structure of the MM, complements the necessity to refer to governance and management concepts tested and backed by standards and methodologies. Moreover, the model advocates continuous learning and improvements in governance in IT outsourcing and good management of outsourced IT services, even when organizations have reached the maximum level (5).

VI. RESEARCH METHODOLOGY

Both ISO 20000 and ISO 38500 standards, and ITIL v3 and COBIT methodologies of best practice in IT management and governance, are a good basis for the study and analysis of governance and management of the outsourced IT services in organizations. That is why they allow the design of a new maturity model that facilitates the achievement of an effective transition to a model of good governance and management of outsourced IT services that, aligned with the core business in organizations, impacts on the effectiveness and efficiency of its management, optimizes its value and minimizes risks.

A questionnaire (survey form) forms the basis of the quantitative study of the maturity model. The questionnaire is based on the attributes or indicators that define the different levels of the model. It contains standard and suitable questions, according to the nature of the research.

Questionnaire responses allow the obtaining or calculation of the level of maturity by applying the scale defined in the model. In addition, questionnaire responses, after being properly analysed, shed light on the current situation of the different organizations studied in governance and management of outsourced IT services.

This research also provides specific case studies carried out at some universities. These case studies put the model into practice in order to draw conclusions. The questions used in the

questionnaire, completed by the universities studied, bring the design of a proposed improvement plan (see Table III) to allow a sequential growth by stages. The growth occurs as a hierarchical progression that should not be reversed, for the aforementioned reasons, and involve a broad range of organizational activities in governance and management of IT outsourcing.

Table III shows one of the five levels (there are five tables, one for each level) of the MM with the key areas or concepts to be improved in order to allow a sequential growth by stages. The first column of the table shows the concepts. The second column of the table shows the objectives to achieve corresponding to the concept in the first column. Finally, the third column shows the actions to accomplish in order to achieve the objectives set in the second column.

TABLE III. IMPROVEMENT PLAN. LEVEL 2

| Level 2 - Initial or improvised | | |
|---|--|--|
| Concept | Improvement Objectives | Improvement Actions |
| Formal Agreement: Contract, agreement or similar (FA) | <ul style="list-style-type: none"> - The formal agreement includes services to be provided, SLA, costs and responsibilities - There are not any clear documented procedures in order to manage outsourced IT services. - There are not any clear processes in order to negotiate with providers | <ul style="list-style-type: none"> - IT Management must understand the necessity that every formal agreement of every outsourced IT service should include the following: services to provide, SLA, costs and responsibilities of the parties |
| Service Measurement (MED) | <ul style="list-style-type: none"> - Informal and reactive measurement (quality, performance and risks) of the IT services provided externally | <ul style="list-style-type: none"> - Measurement of quality, performance and risks of outsourced IT services is essential to meet the expectations and business needs. In addition, measurements help detect early potential problems. Therefore, it would be necessary to carry out this measurement, even if it is informal and reactive |
| Monitoring and adjustments of outsourcing (MON) | <ul style="list-style-type: none"> - Informal supervision of outsourced IT services, the associated risks and the provision of services - The process and the indicators used are not optimized - Indicators hardly affect penalties, contracting and negotiation of outsourced IT services | <ul style="list-style-type: none"> - Supervision of external service providers is essential for monitoring. It would be advisable to carry out the supervision of outsourced IT services, the associated risks and the provision of services, even if the supervision is informal - Organization must understand the necessity to make a process to supervise outsourced IT services, the provision of services and associated risks. In order to accomplish that, it is needed to use Key Performance Indicators (KPIs) and Key Goals Indicators (KGIs). Both process and indicators will be optimized as time goes by. Therefore, they are not optimized yet. - The results of KPIs and KGIs should begin to indicate the degree of compliance with the agreements signed. This fact should be quantified and organizations should turn it into sanctions or penalties based on the level of non-compliance with established agreements. Besides, IT governance members should start to be aware of the possibility of adjusting the process of procurement and monitoring of outsourced IT services, based on the results of KPIs and KGIs |
| Alignment IT-Business (ALI) | <ul style="list-style-type: none"> - The requirements of the outsourced IT services are hardly defined, implemented and aligned with business objectives | <ul style="list-style-type: none"> - Managers should begin to understand that it is necessary to define a clear IT strategy based on the business needs of the organization. Managers should begin to design an IT strategic plan in line with the organization, in order to achieve the objectives. Thus, alignment IT business is easier. A good start to achieve the objectives would be to define and implement gradually the requirements of the outsourced IT services in order to keep them aligned with the business objectives of organizations |
| IT Governance Structure (EOG) | <ul style="list-style-type: none"> - Organization starts to build an organizational structure of IT Government where the CIO or equivalent is the backbone and member of the Board of Directors. This structure consists of at least: IT Strategy Committee, IT Steering Committee, Projects Office and Services | <ul style="list-style-type: none"> - The government team or the management should take responsibility to start creating a structure of decision making related to IT, where the CIO should be the protagonist and the backbone and integrator of IT strategy in the organization. Therefore, CIO should be part of the Board of Directors. Also, it would be necessary to establish at least the following: |

| | | |
|---------------------------------------|--|---|
| | Commission | <ul style="list-style-type: none"> - <u>IT Strategy Committee</u>. It should design the strategy and high-level policies related to IT of the university. It should be composed of all managers with IT strategic responsibility , besides CIO that should be part of; - <u>IT Steering Committee</u>. It should design and implement IT projects in order to meet strategic planning designed by the IT Strategy Committee; - <u>IT Projects Office</u>. It should manage IT projects designed by the IT Steering Committee in order to meet strategic planning designed by the IT Strategy Committee; and - <u>Services Commission</u>. It should have a composition to represent all end users of IT services |
| Service Level Agreement (SLA) | <ul style="list-style-type: none"> - There is an SLA for every outsourced IT service - SLA includes: the essential aspects of the outsourced IT service, such as the service description and the service availability; and the responsibilities of the parties | <ul style="list-style-type: none"> - SLA is the reference document where is stated how the service signed between the service provider and the customer is provided. It would be necessary to have an SLA for every outsourced IT service - Every outsourced IT service should include the following: description of the service; availability; and responsibilities of the parties |
| IT Service Registration (RSS) | <ul style="list-style-type: none"> - Service catalog clearly defined and updated | <ul style="list-style-type: none"> - A service catalog clearly defined and updated should be created with all current active services. The following aspects might be missing in the service catalog: the conditions of provision of services; SLA; costs; and mutual responsibilities of the parties |
| Incident and Problem Management (GIP) | <ul style="list-style-type: none"> - Incident Management Process (GI) is implemented - The degree of optimization of incident management process is good - Tools that manage incidents are optimized and they allow registration, tracking and monitoring of incidents - Barely there is link between incident management and service level management | <ul style="list-style-type: none"> - The main objectives of the incident management process are: to detect any change in services; recording and classifying these changes; and assign personnel to restore the service. It would be necessary to have this process implemented with a good degree of optimization. This responsibility would fall into the IT director - The tools for recording, tracking and monitoring incidents, should be optimized, because they constitute an important component of the incident management - The head of the IT department should start being aware of the importance of the link between incident management and service level management |
| Testing and Deployment (PYD) | <ul style="list-style-type: none"> - Professional IT staff. There is no improvisation at all | <ul style="list-style-type: none"> - The deployment and testing of IT services are essential for the proper functioning of services when they are in production. Improvisation should stay out of IT operations when deploying and testing IT services. Also, it is necessary that the IT staff (internal and external) and end users are well trained |
| Legislation (LEG) | <ul style="list-style-type: none"> - Possible loopholes in data protection, data processing, location where data processing takes place, clauses for the transfer of data and standard contractual clauses for the transfer of personal data to third countries, have been corrected | <ul style="list-style-type: none"> - IT service providers access or treat personal data of employees of organizations that hire their services. In the case of universities, IT service providers access to personal data of teaching, research and administrative personnel, besides the personal data of students who have or have had any type of registration at the universities. Therefore, in order to avoid the risks in the privacy of individuals caused by the processing of personal data carried out by third parties, it would be necessary to correct the possible loopholes that might exist in data protection, data processing, location where data processing takes place, clauses for the transfer of data and standard contractual clauses for the transfer of personal data to third countries |

VII. APPLICABILITY OF THE MODEL

In order to implement and evaluate the applicability of the model and the assessment instrument designed, by following the measurement process through the improvement plan, we need to apply the model in institutions of higher education. The purpose of this research is not to obtain statistical results of the universe of the study, which in our case are all public and private institutions of higher education in Spain, but to evaluate

the usefulness of the proposed model through several case studies. Samples from a universe of study can be classified into probabilistic and non-probabilistic. In the former, the key feature is that every element of the universe has a certain probability of entering the sample, and this probability can be calculated mathematically and accurately. In the latter, the opposite occurs and the researcher doesn't know the error that can be introduced in the assessments.

There are several types of samples in non-probabilistic samples. One of them is the intentional sample. In this kind of sample, units are chosen arbitrarily according to the characteristics of the sample that the researcher finds relevant. Therefore, knowledge and personal opinions are used to identify those units to be included in the sample. It is based mainly on the experience of someone with the population. These samples are very useful and are often used in case studies.

The case studies in this research have been conducted on an intentional sample, under the schema of positivist research, one of the three approaches there are in qualitative research. This kind of research assumes that reality is given objectively and it can be described by measurable properties (characteristics or indicators of the model) that are independent of the observer and the instruments used. Positivist studies try to test the theory in an attempt to increase the predictive understanding of a phenomenon. In line with this Orlikowski and Baroudi [15] labeled research in information systems as positivist if there was evidence of formal propositions, quantifiable measures of variables, tests of hypothesis, and draw conclusions about a phenomenon from a sample of a population estimated.

Therefore, the case study undertaken to implement the applicability of the model, have been performed on an intentional non-probabilistic sample and under the scheme of the positivist qualitative research.

In the case study, we will apply the established scales, which will rate the university surveyed and the object of study, at a level of maturity within the MM. Depending on the level of maturity in which the university is rated, improvement actions, according to the improvement plan, will be proposed to achieve a target level.

The measurement process to ascend in the MM is as follows (see Fig. 1):

- 1) Perform an initial measurement after completing the questionnaire;
- 2) Set goals (benchmark);
- 3) Identify the gaps between the current measurement and the objectives set;
- 4) Recommend actions and policies to be implemented within the improvement plan to ascend in the MM; and
- 5) Once corrective actions have been implemented, perform a new measurement.

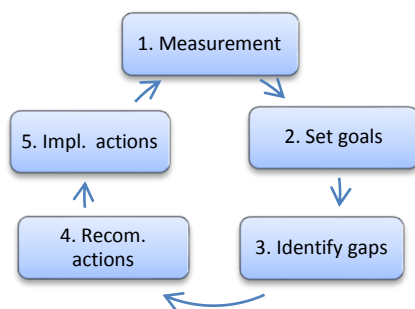


Fig. 1. Measurement process

The continuous improvement plan to apply the established scales (see Table II) in order to achieve a target level is as follows (see Fig. 2):

1) Initial measurement of the current level of the institution studied after completing the questionnaire. Equivalent to step 1 of the measurement process;

2) Identify improvement objectives using the values of the indicators. Equivalent to steps 2 and 3 of the measurement process; and

3) Implement improvement actions or practices in order to achieve the improvement objectives identified in stage 2. Equivalent to steps 4 and 5 of the measurement process. Back to the first stage.

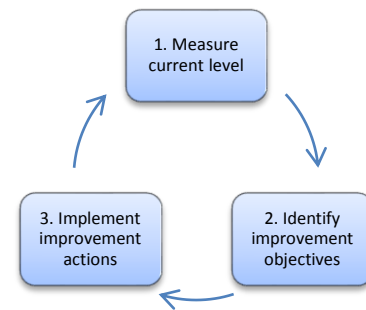


Fig. 2. Continuous improvement plan

VIII. RESULTS

A. A case study

In order to accomplish the applicability of the model by case studies, we selected three institutions as an intentional representative sample, and we sent an email to the IT managers asking them to participate in this research project. The email contained a link to the questionnaire and the recipients were asked to complete the questionnaire hosted on the www.encuestafacil.com surveys website. This website provides a tool to design questionnaires in an intuitive and effective way. Also, this web tool allows to export the results / responses of questionnaires to a spreadsheet.

After collecting the responses of the questionnaire, completed by the three institutions selected through the web-based questionnaire and using the tool hosted on www.encuestafacil.com, we start the first stage of the continuous improvement plan (see Fig. 2). In this case, we are going to use only the responses of one institution that we keep anonymous in order to protect its privacy and confidentiality.

1) Stage 1

Once the questionnaire is completed by the institution and the responses are collected by us, we must measure the current level of the institution studied. Table IV shows the responses of the questionnaire. The first column of the table shows the level. The second column shows the indicators corresponding to the level in the first column. Finally, the third column shows the responses corresponding to the indicators in the second column.

TABLE IV. RESPONSES OF THE QUESTIONNAIRE

| LEVEL | INDICATOR | HEI-1 |
|--------------|--------------|-------|
| 1 | ALI3-(1)2345 | Y |
| | LEG1-(1)23 | Y |
| | Maturity 1 | 1 |
| 2 | ACF2b | Y |
| | ACF2d | Y |
| | ACF2e | Y |
| | ACF2g | Y |
| | MON3-(2)34 | Y |
| | MON4-(2)34 | Y |
| | ALI3-1(2)345 | Y |
| | EOG1 | N |
| | EOG2 | N |
| | EOG4 | Y |
| | EOG6 | Y |
| | EOG7 | N |
| | SLA1 | Y |
| | SLA2a | Y |
| | SLA2d | Y |
| | RSS1 | Y |
| | GIP1 | Y |
| | GIP2-(2)3 | Y |
| | GIP3-(2)345 | Y |
| | PYD1 | Y |
| | LEG1-1(2)3 | Y |
| | Maturity 2 | - |
| | 3 | ACF 1 |
| ACF2a | | Y |
| ACF2c | | Y |
| ACF2f | | Y |
| ACF3a | | Y |
| ACF3b | | Y |
| ACF3c | | Y |
| ACF5 | | Y |
| ACF6-(3)45 | | Y |
| MED1 | | N |
| MON1 | | Y |
| MON2 | | NR/DK |
| MON3-2(3)4 | | N |
| MON4-2(3)4 | | N |
| MON5a | | Y |
| MON5b | | Y |
| ALI3-12(3)45 | | Y |
| EOG3 | | N |
| EOG5 | | Y |
| SLA2c | | Y |
| SLA2e | | Y |
| SLA3 | | N |
| GIP2-2(3) | | Y |
| GIP3-2(3)45 | | Y |
| GIP4 | | Y |
| GIP5-(3)4 | | Y |
| CAM1 | | Y |
| RIN3 | | NR/DK |
| LEG1-12(3) | | N |
| PAS1a | | Y |
| PAS1b | | Y |
| PAS1c | Y | |
| PAS1d | Y | |
| PAS1e | Y | |

| | | |
|------------|--------------|-------|
| 4 | PAS1f | Y |
| | Maturity 3 | - |
| | ACF4 | N |
| | ACF6-3(4)5 | N |
| | MON3-23(4) | N |
| | MON4-23(4) | N |
| | MON6 | Y |
| | ALI1 | N |
| | ALI2 | NR/DK |
| | ALI3-123(4)5 | Y |
| | SLA2f | N |
| | SLA2g | Y |
| | GIP3-23(4)5 | Y |
| | GIP5-3(4) | N |
| | CAM2 | Y |
| | CAM3 | Y |
| | RIN1 | NR/DK |
| | RIN2 | |
| | RIN4 | Y |
| | GDC1 | NR/DK |
| | GDC2 | NR/DK |
| | GDC3-(4)5 | |
| | GAF1 | Y |
| GCO1 | NR/DK | |
| Maturity 4 | - | |
| 5 | ACF6-34(5) | N |
| | GC1 | NR/DK |
| | ALI3-1234(5) | N |
| | SLA2b | N |
| | RSS2a | Y |
| | RSS2b | N |
| | RSS2c | N |
| | RSS2d | N |
| | GIP3-234(5) | N |
| | CAM4 | Y |
| | CPE1 | NR/DK |
| | CPE2 | NR/DK |
| | CGF1 | N |
| | GDC3-4(5) | |
| | GAF2 | NR/DK |
| | GAF3 | |
| | GCO2 | |
| Maturity 5 | - | |

Table IV shows indicators and responses needed to qualify for a maturity level in the institution selected. In order to qualify for a maturity level, all indicators of that level must be responded positively ('Y' in the third column of the table IV). Therefore, the institution selected is at level 1.

2) Stage 2

At this stage, improvement objectives must be identified in order to enable and facilitate studied institution to move up the maturity model. But first, we must set goals to achieve, which in our case is to consolidate the level 2, and afterwards identify differences between the objectives to achieve and the current assessment. Indicators of level 2 with the value to 'N' (see Table IV) allow identifying improvement objectives. Red cells in Table V identify indicators that allow to identify improvement objectives.

TABLE V. INDICATORS THAT ALLOW IDENTIFYING IMPROVEMENT OBJECTIVES

| LEVEL | INDICATOR | HEI-1 |
|------------|--------------|-------|
| 1 | ALI3-(1)2345 | S |
| | LEG1-(1)23 | S |
| | Maturity 1 | 1 |
| 2 | ACF2b | S |
| | ACF2d | S |
| | ACF2e | S |
| | ACF2g | S |
| | MON3-(2)34 | S |
| | MON4-(2)34 | S |
| | ALI3-1(2)345 | S |
| | EOG1 | N |
| | EOG2 | N |
| | EOG4 | S |
| | EOG6 | S |
| | EOG7 | N |
| | SLA1 | S |
| | SLA2a | S |
| | SLA2d | S |
| | RSS1 | S |
| | GIP1 | S |
| | GIP2-(2)3 | S |
| | GIP3-(2)345 | S |
| | PYD1 | S |
| LEG1-1(2)3 | S | |
| Maturity 2 | - | |

The indicators that allow to identify improvement objectives in the institution studied are: EOG1; EOG2; and EOG7.

Improvement objectives must be searched in the tables of improvement objectives and improvement actions by levels (see Table III). In the case of the institution studied, it should move up second level in order to move to a higher level gradually, consolidating steps. Thus, we have to look for the improvement objectives of level 2 in Table III. Firstly, we must find out the concepts, located in the first column of the Table III, corresponding to the indicators selected in the stage 2 of continuous improvement plan. Finally, we need the table of metrics and questionnaire (see Table II) in order to identify the characteristics that define the indicators selected.

Therefore, the improvement objectives are the following:

- Organization starts to build an organizational structure of IT Government where the CIO or equivalent is the backbone and member of the Board of Directors [EOG1].
- This IT structure has an IT Strategy Committee [EOG2] and a Services Commission [EOG7].

3) Stage 3

In the last stage of continuous improvement plan we must identify and implement the actions corresponding to the improvement objectives identified in the previous stage (stage 2). Implementation of these actions successfully will allow the institution studied to move up level 2 of the maturity model.

Therefore, the improvement actions to implement corresponding to the improvement objectives identified in stage 2, are the following:

- The government team or the management should take responsibility to start creating a structure of decision making related to IT, where the CIO should be the protagonist, the backbone and integrator of IT strategy in the organization. In addition, CIO should be part of the Board of Directors. Also, it would be necessary to establish the following:
 - IT Strategy Committee. It should design the strategy and high-level policies related to IT of the university. It should be composed of all managers with IT strategic responsibility, besides CIO that should be as well part of.
 - Services Commission. It should have a composition to represent all end users of IT services.

Once improvement objectives and improvement actions have been identified, it is time to implement improvement actions. Once these actions have been implemented successfully, the third stage of the continuous improvement plan has been finished. That means institution studied has achieved the goals set in stage 2 of the continuous improvement plan. In other words, institution studied has reached level 2 of the maturity model.

The continuous improvement plan is designed in such a way that allows moving up gradually in the model by repeating the three stages as many times as necessary. Therefore, the next step would be to perform the first stage of the plan in order to measure the institution again. This first step is critical because the time between two measurements could be long, and some key practices done in the previous measurement, could not be done in the following measurement; and on the contrary, some key practices not done in the previous measurement, could be done in the next measurement, in addition to the practices to implement identified in stage 2 and implemented in stage 3 of the previous cycle of the continuous improvement plan.

IX. CONCLUSIONS

In order to design the proposed innovative maturity model, we studied in detail every reference on the provision of IT services that there is in the ISO 20000 and ISO 38500 standards and ITIL v3 and COBIT methodologies. In addition, we investigated the relevant literature and failed to find any maturity model that brings together the previous methodologies with a specific focus on IT outsourcing. As a result, a number of concepts and subconcepts were categorized to form the basis of the maturity model.

Furthermore, models, standards and guidelines are recommended in order to enable and facilitate adaptation to universities so that they can move up the maturity model. Thus, the model, based on standards and best practices, is designed to achieve excellence in the management of IT outsourcing. The applicability of the study, allows universities to meet the goal of effective transition to a model of good governance and good management of outsourced IT services. Aligned with the core business of universities (education, research and innovation) this will impact on the effectiveness and efficiency of their management, optimize value and minimize risks.

The model allows organizations under study ascend from one level of maturity to the next without skipping any intermediate level. In practice, organizations can accomplish specific practices in upper levels. However, this does not mean they can skip levels, since optimum results are unlikely if practices in lower levels go unfulfilled.

MM advocates continuous learning and improvements in governance and management of outsourced IT services, even when institutions have consolidated the highest level of the maturity model (level 5).

This study recognizes that it is unlikely to achieve maximum effectiveness and efficiency in the government and management of outsourced IT services, in a higher education institution in a relatively short period of time. The structure of the model proposed, organized in levels, provides a general understanding of the gradual and holistic development of IT governance and management of outsourced IT services. MM expects to be an effective diagnostic tool to measure the efforts made around IT outsourcing in higher education institutions, in addition to a coherent roadmap to guide institutions in their efforts to provide their teaching, research, and administrative staff, and ultimately their students, with a quality and effective IT services in line with the digital technological era of the XXI century.

On the basis of this research, by designing an assessment tool along with the maturity model that allows independent validation and practical application of the model, this study will allow higher education institutions to achieve an effective transition to a model of good governance and management of outsourced IT services in order to meet successfully the requirements demanded in this complex internet age.

REFERENCES

- [1] O. Adalakun, IT Outsourcing Maturity Model, in: Proceedings of the 12 European conference on Information System, (ECIS) Turku Finland, June 14-16, 2004
- [2] E. Beulen and P. Ribbers, Managing complex IT outsourcing-partnerships, in: Proceedings of the 35th Annual Hawaii International Conference on System Sciences, 7-10 Jan 2002, vol., no., pp.10 pp., doi: 10.1109/HICSS.2002.994424
- [3] T. Dahlberg and P. Lahdelma, IT Governance Maturity and IT Outsourcing Degree: An Exploratory Study, in: the 40th Annual Hawaii International Conference on System Sciences, Jan. 2007, vol., no., pp.236a,236a, doi: 10.1109/HICSS.2007.306
- [4] A.M. Fairchild, Information technology outsourcing (ITO) governance: an examination of the outsourcing management maturity model, in: Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 5-8 Jan. 2004, vol., no., pp.8 pp., doi: 10.1109/HICSS.2004.1265565
- [5] Fan Jing Meng, Xiao Yang He, Yang, S.X. and Peng Ji, A Unified Framework for Outsourcing Governance, in: the 9th IEEE International Conference on E-Commerce Technology, and in: the 4th IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services, CEC/EEE 2007, vol., no., pp. 367, 374, doi: 10.1109/CEC-EEE.2007.16
- [6] E. Fernández, UNiTIL: Gobierno y Gestión del TIC basado en ITIL, in III Congreso Interacadémico 2008 / itSMF España. Last Access on 5th Sept 2011at http://www.uc3m.es/portal/page/portal/congresos_jornadas/congreso_itsmf/UNiTIL%20Gobierno%20y%20Gestion%20de%20TIC%20basado%20en%20ITIL.pdf
- [7] P. Gottschalk, and H. Solli-Saether, Critical success factors from IT outsourcing theories: an empirical study, *Industrial Management & Data Systems*, Vol. 105 Nos 5/6, pp. 685-702, 2005
- [8] P. Gottschalk and H. Solli-Sæther, Maturity model for IT outsourcing relationships, *Industrial Management & Data Systems*, Vol. 106 Iss: 2, pp.200 – 212, 2006
- [9] T.P. Herz, F. Hamel, F. Uebernickel and W. Brenner, Towards a Multisourcing Maturity Model as an Instrument of IT Governance at a Multinational Enterprise, in: the 44th Hawaii International Conference on System Sciences, Jan. 2011, vol., no., pp.1,10, 4-7, doi: 10.1109/HICSS.2011.448
- [10] ITGI. Last Access on 26th February 2014 at <http://www.itgi.org/>
- [11] K. Lyytinen, Penetration of information technology in organizations, *Scandinavian Journal of Information Systems*, 3, 87–109, 1991
- [12] R. Mobarhan, A.A. Rahman and M. Majidi, Outsourcing management framework based on ITIL v3 framework, in: the 7th International Conference on Information Technology in Asia (CITA 11), July 2011, vol., no., pp.1,5, 12-13, doi: 10.1109/CITA.2011.5999536
- [13] Ning Su and N. Levina, Global Multisourcing Strategy: Integrating Learning From Manufacturing Into IT Service Outsourcing, *IEEE Transactions on Engineering Management*, Nov. 2011, vol.58, no.4, pp.717,729, doi: 10.1109/TEM.2010.2090733
- [14] R.L. Nolan, Managing the computer resource: a stage hypothesis, *Communications of the ACM*, 16 (7). 399, 1973, doi: [10.1145/362280.362284](https://doi.org/10.1145/362280.362284)
- [15] W.J. Orlikowski and J.J. Baroudi, Studying Information Technology in Organizations: Research Approaches and Assumptions, *Information Systems Research* (2) 1991, pp. 1-28
- [16] Nandish Patel, Emergent Forms of IT Governance to Support Global eBusiness Models, *Journal of Information Technology Theory and Application (JITTA)*, 4 (2), Article 5, 2002, Last access on 19th January 2014 at <http://aisel.aisnet.org/jitta/vol4/iss2/5>
- [17] C. Schwarz, Toward an understanding of the nature and conceptualization of outsourcing success, *Information & Management*, 51(1), 2014, pp. 152-164
- [18] Thompson S.H. Teo and Anol Bhattacharjee, Knowledge transfer and utilization in IT outsourcing partnerships: A preliminary model of antecedents and outcomes, *Information & Management*, 51(2), 2014, pp. 177-186
- [19] R. Yanosky and J. Borreson Caruso, Process and Politics: IT Governance in Higher Education, ECAR Key Findings. EDUCASE, 2008, Last Access on 6th March 2013 at <http://net.educause.edu/ir/library/pdf/ekf/EKF0805.pdf>
- [20] R. Yanosky and J. McCredie, IT Governance: Solid Structures and Practical Politics, ECAR Symposium, Boca Ratón, Florida, 2007

Modification of CFCM in The Presence of Heavy AWGN for Bayesian Blind Channel Equalizer

Changkyu Kim

Dept. of Information & Communication Eng.
Donggeui University
Pusan, Korea 614-714

Soowhan Han

Dept. of Multimedia Engineering
Donggeui University
Pusan, Korea 614-714

Abstract—In this paper, the modification of conditional Fuzzy C-Means (CFCM) aimed at estimation of unknown desired channel states is accomplished for Bayesian blind channel equalizer under the presence of heavy additive Gaussian noise (AWGN). For the modification of CFCM to search the optimal channel states of a heavy noise-corrupted communication channel, a Gaussian weighted partition matrix, along with the Bayesian likelihood fitness function and the conditional constraint of ordinary CFCM, is developed and exploited. In the experiments, binary signals are generated at random and transmitted through both types of linear and nonlinear channels which are corrupted with various degrees of AWGN, and the modified CFCM estimates the channel states of those unknown channels. The simulation results, including the comparison with the previously developed algorithm exploiting the ordinary CFCM, demonstrate the effectiveness of proposed modification in terms of accuracy and speed, especially under the presence of heavy AWGN. Therefore, the proposed modification can possibly constitute a search algorithm of optimal channel states for Bayesian blind channel equalizer in severe noise-corrupted communication environments.

Keywords—Gaussian Partition Matrix; Conditional Fuzzy C-Means; Channel States; Bayesian Blind Equalizer

I. INTRODUCTION

Channel equalization is a major issue in digital communications, because a channel is easily affected by inter-symbol-interference (ISI) with both linear and nonlinear distortions in the presence of AWGN. The task of channel equalization is to minimize those distortions to recover the transmitted sequence. In general, there exist two kinds of equalizers in digital communication systems: data aided (trained) equalizers and blind equalizers. For trained equalizers, a reference signal is required, increasing the bandwidth. However, in blind equalizations, the original transmitted message is recovered only from the received sequence that is corrupted by noise without any training sequence or a priori knowledge of the channel. As a result, the use of blind equalizers is preferred in high-speed communication systems to reduce ISI without increasing overhead costs [1][2]. Because of inherent simplicity, most available blind equalization algorithms focus on linear channels that are often inadequate for modeling channels which exhibit nontrivial nonlinearities [3]. In practical world, the equalization of nonlinear is often required such as in high power amplifiers as well as in high-density magnetic and optical storage channels. Therefore, the blind equalization

method handled in this paper must be dealt with both linear and nonlinear channels, which is independent of the type of channel structure.

Traditionally, channel equalization has been considered equivalent to inverse filtering. The optimal solution, based on maximum likelihood sequence estimation (MLSE) [4], has a complexity that grows exponentially with the dimension of the channel impulsive response (Viterbi algorithm). Alternatively, several nonlinear detection procedures have been proposed to address this problem with varying degrees of success, such as multi-layered perceptrons (MLPs) [5], radial basis function networks (RBFNs) [6], recurrent RBFNs [7], self-organizing feature maps (SOFMs) [8][9], wavelet neural networks [10], kernel Adeline (KA) [11], support vector machines (SVMs) [12] and Genetic Algorithms[13][14]. Such structures usually outperform linear equalizers, especially when non-minimum phase channels are encountered. They can also compensate for nonlinearities in the channel. However, they still suffer from the relatively high computational cost such as the iterative reweighted quadratic procedure of SV in [12]. The simplex Genetic Algorithm (GA) in [13] estimates the optimal channel output states instead of estimating the channel parameters in a direct manner. The desired channel states of an unknown channel were constructed from these estimated channel output states, and placed at the center of RBF equalizer. With this approach, the complex modeling of the nonlinear channel can be avoided and the method works well within a simple single input single output (SISO) communication environment. Additionally, this kind of approach can be applied to a linear channel as well, because it does not estimate the channel parameters but the channel output states directly, which is not dependent on the type of the channel structure. However, the GA based algorithms may visibly suffer from their poor convergence properties. Recently, to overcome this weakness, Fuzzy C-Means (FCM), one of the representative clustering algorithms which exhibits shorter processing time than the GA-based methods, has been modified and applied, and the faster convergence speed along with the reliable estimation accuracy in search of the optimal channel output states has been achieved [15][16]. Especially, the algorithm based on CFCM clearly outperforms the GA and FCM approaches in terms of speed and accuracy [16]. The CFCM was first introduced in [17], and successfully applied to channel equalization problem [16][18]. The conditioning aspect of CFCM, which describes a level of involvement of incoming input pattern in the constructed clusters, influences the clustering mechanism and improves the estimation accuracy

of an unknown channel states for blind channel equalization. However, in the presence of heavy AWGN that often arises in a high speed communication channel, the estimation accuracy of CFCM presented in [16] needs the higher level of reliability, even though it is superior to other FCM or GA-related algorithms. This leads to the consideration of the modification of CFCM clustering mechanism, which makes it more robust to the heavy noise. In this study, the modification is accomplished by using a Gaussian weighted partition matrix during the clustering procedure of CFCM. The use of Gaussian weights for a partition matrix instead of ordinary Euclidean distance measuring can help to search the correct channel states of an unknown channel, because the received sequence under the presence of AWGN is a scattered random process having conditional Gaussian density functions centered at each of the desired channel states. More details on this modification of CFCM are explained in Section 5. Before that, an optimal Bayesian equalizer for a linear/nonlinear channel is introduced in the next section and the reconfiguration procedure of desired channel states with channel output states is discussed in Section 3. In Section 4, the fitness function for the proposed CFCM is derived. This study is an extension of previous work [16] and thus the similarity of the structure of Section 2 and 3 can be found in [16]. Finally, the simulation results including some comparative studies with early work [16] and conclusions are provided in Section 6 and 7, respectively.

II. OPTIMAL BAYESIAN EQUALIZER FOR A LINEAR/NONLINEAR CHANNEL

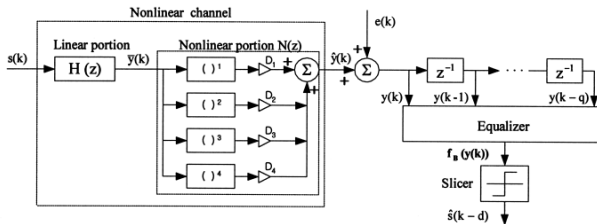


Fig. 1. Channel equalization with ISI

A general channel equalization system with ISI is illustrated in Fig. 1. The digital information symbol $s(k)$, which is assumed to be an equiprobable and independent binary sequence taking values from a two-valued set $\{\pm 1\}$, is transmitted through a nonlinear dispersive channel. Here the nonlinear channel is composed of a linear part with transfer function $H(z)$, whose output $\bar{y}(k)$ described by (1), and a nonlinear component $N(z)$, whose output $\hat{y}(k)$ governed by (2).

$$\bar{y}(k) = \sum_{i=0}^p h(i)s(k-i) \quad (1)$$

$$\hat{y}(k) = D_1\bar{y}(k) + D_2\bar{y}(k)^2 + D_3\bar{y}(k)^3 + D_4\bar{y}(k)^4 \quad (2)$$

In (1) and (2), p is the channel order and D_i stands for the coefficient of the i^{th} nonlinear term which is possibly caused by nonlinearities associated with nonlinear devices used in the transmitter and the receiver. For a linear channel model, the nonlinear terms are set to $D_1=1$, $D_2=0$, $D_3=0$ and $D_4=0$. The

noise-free observation vector, $\hat{y}(k)$ expressed by (3), is referred to as the desired channel states, and for a specific equalizer order denoted by q , there exist $M = 2^{p+q+1}$ different patterns.

$$\hat{y}(k) = [\hat{y}(k), \hat{y}(k-1), \dots, \hat{y}(k-q)] \quad (3)$$

These M desired channel states can be partitioned into two sets, $\mathbf{Y}_{q,d}^{+1}$ and $\mathbf{Y}_{q,d}^{-1}$, as shown in (4) and (5), depending on the value of $s(k-d)$, where d is the required time delay.

$$\mathbf{Y}_{q,d}^{+1} = \{ \hat{y}(k) | s(k-d) = +1 \} \quad (4)$$

$$\mathbf{Y}_{q,d}^{-1} = \{ \hat{y}(k) | s(k-d) = -1 \} \quad (5)$$

Finally, the desired channel states, $\hat{y}(k)$, are corrupted by the AWGN, $e(k)$, and thus the channel observation vector (input of equalizer) $\mathbf{y}(k)$ can be described as

$$\mathbf{y}(k) = \hat{y}(k) + e(k) \quad (6)$$

The task of the equalizer is to produce the estimated sample $\hat{s}(k-d)$ which has the same value as the transmitted symbol $s(k-d)$, based on the noise-corrupted observation vector, $\mathbf{y}(k)$. Because of the AWGN, the observation vector, $\mathbf{y}(k)$, is a random process having conditional Gaussian density functions centered at each of the desired channel states, $\hat{y}(k)$. The determination of the value of $s(k-d)$ becomes a decision problem. The optimal symbol-by-symbol spaced equalizer decision function is provided by the maximum a-posteriori probability criteria and is called Bayesian equalizer. The decision function for Bayesian equalizer [19] can be represented as follows,

$$f_B(\mathbf{y}(k)) = \sum_{i=1}^{n_s^{+1}} \exp(-\|\mathbf{y}(k) - \mathbf{y}_i^{+1}\|^2 / 2\sigma_e^2) - \sum_{i=1}^{n_s^{-1}} \exp(-\|\mathbf{y}(k) - \mathbf{y}_i^{-1}\|^2 / 2\sigma_e^2) \quad (7)$$

$$\hat{s}(k-d) = \text{sgn}(f_B(\mathbf{y}(k))) = \begin{cases} +1, & f_B(\mathbf{y}(k)) \geq 0 \\ -1, & f_B(\mathbf{y}(k)) < 0 \end{cases} \quad (8)$$

where \mathbf{y}_i^{+1} and \mathbf{y}_i^{-1} are the desired channel states belonging to sets $\mathbf{Y}_{q,d}^{+1}$ and $\mathbf{Y}_{q,d}^{-1}$, respectively, and their number of elements in these sets are denoted by n_s^{+1} and n_s^{-1} . Furthermore σ_e^2 is the noise variance. From (7) and (8), the evaluation of desired channel states is essential for the optimal Bayesian equalizer, and the performance of Bayesian blind equalizer highly depends on the correct estimation of the desired channel states, \mathbf{y}_i^{+1} and \mathbf{y}_i^{-1} , only from the noise-corrupted observation vector, $\mathbf{y}(k)$. In this study, the modification of CFCM with Gaussian weighted partition matrix is presented to search the optimal states of an unknown channel under the presence of heavy AWGN. After the estimation of the desired channel states, the equalization for the reconstruction of the transmitted symbols is straightforward with (7) and (8).

III. RECONFIGURATION OF DESIRED CHANNEL STATES WITH CHANNEL OUTPUT STATES

The knowledge of the desired channel states, \mathbf{y}_i^{+1} and \mathbf{y}_i^{-1} , is essential for the Bayesian equalizer. If the channel order is

taken as $p=1$ with transfer function $H(z) = 0.5 + z^{-1}$, the equalizer order q is equal to 1, the time delay d is also set to 1, and the nonlinear portion is described by $D_1 = 1, D_2 = 0.0, D_3 = -0.9, D_4 = 0.0$ (see Fig. 1), then the eight different desired channel states ($2^{p+q+1} = 8$) may be observed at the receiver in a noise-free case. The input sequences, the desired channel states and the output of the equalizer for this channel are shown in Table 1. From this table, it can be seen that the values of the desired channel states $[\hat{y}(k), \hat{y}(k-1)]$ are composed of the elements of the scalar channel states called "channel output states", $\{a_1, a_2, a_3, a_4\}$, where for this particular channel they have $a_1 = -1.5375, a_2 = -0.3875, a_3 = 0.3875$ and $a_4 = 1.5375$. The only difference between the desired channel states and channel output states is that the first are vectors while the latter are scalars. The length of this scalar dataset, \tilde{n} , is determined by the channel order, p , such as $2^{p+1} = 4$, which is independent of the equalizer order. As shown in Table 1, the desired channel states for $\mathbf{Y}_{1,1}^{+1}$ (positive states) and $\mathbf{Y}_{1,1}^{-1}$ (negative states) are $(a_1, a_1), (a_1, a_2), (a_3, a_1), (a_3, a_2)$ and $(a_2, a_3), (a_2, a_4), (a_4, a_3), (a_4, a_4)$, respectively. A change in the decision delay only changes some of the positive states to negative and the equal number of the negative states to positive. Additionally, it can be applied for a linear model as well, where nonlinear terms of channel, D_2, D_3 , and D_4 , are equal to zero. In case of the linear model, the elements of data set $\{a_1, a_2, a_3, a_4\}$ become $1.5, -0.5, 0.5$ and -1.5 , respectively, and are shown in Table 1 by parentheses. The desired channel states of these nonlinear and linear models are illustrated in Fig. 2. This relationship of desired channel states and channel output states is always valid for the channel that has a one-to-

one mapping between the channel inputs and outputs [13] and is successfully used in [14]-[16]. Additionally, it can be easily extended with a higher channel order such as $p=2$, which is evaluated in the experimental section. If the channel order p is 2 with $H(z) = 0.3482 + 0.8704z^{-1} + 0.3482z^{-2}$, there exist the sixteen desired channel states ($2^{p+q+1} = 16$) composed of the eight channel output states ($\tilde{n} = 2^{p+1} = 8, a_1, a_2, a_3, \dots, a_8$). The desired channel states, $(a_1, a_1), (a_1, a_2), (a_2, a_3), (a_2, a_4), (a_5, a_1), (a_5, a_2), (a_6, a_3), (a_6, a_4)$, belong to $\mathbf{Y}_{1,1}^{+1}$, and $(a_3, a_5), (a_3, a_6), (a_4, a_7), (a_4, a_8), (a_7, a_5), (a_7, a_6), (a_8, a_7), (a_8, a_8)$ belong to $\mathbf{Y}_{1,1}^{-1}$, where $a_1, a_2, a_3, \dots, a_8$ are $2.0578, 1.0219, -0.1679, -0.7189, 1.0219, 0.1801, -0.7189$ and -1.0758 , respectively. This channel can be found in [16] as well. As shown in Table 1, the desired channel states for both types of linear and nonlinear can be constructed with the channel output states if channel order, p , is assumed to be known, and thus the main problem of blind equalization moves its focus onto the determination of the optimal channel output states only from the received patterns.

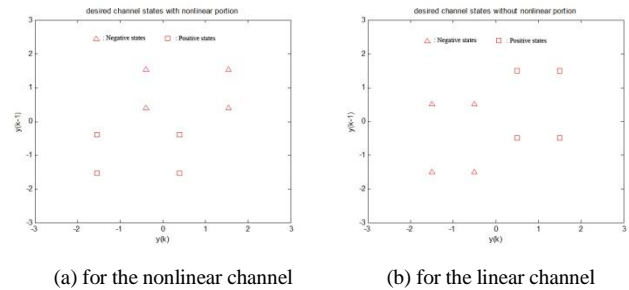


Fig. 2. Desired channel states (noise-free) for the nonlinear and linear channels shown in Table 1 (positive(□) and negative(△) states).

TABLE I. RECONFIGURATION OF DESIRED CHANNEL STATES BY CHANNEL OUTPUT STATES FOR NONLINEAR AND LINEAR MODELS

| Nonlinear channel with $H(z) = 0.5 + 1.0z^{-1}, D_1 = 1, D_2 = 0.0, D_3 = -0.9, D_4 = 0.0$, and $d=1$ | | | |
|--|---|--|---------------------|
| Linear channel with $H(z) = 0.5 + 1.0z^{-1}, D_1 = 1, D_2 = 0, D_3 = 0, D_4 = 0$, and $d=1$ | | | |
| Input sequences | Desired channel states | | Output of equalizer |
| $s(k) s(k-1) s(k-2)$ | $\hat{y}(k)$ () used for the lineal model | By channel output states, $\{a_1, a_2, a_3, a_4\}$ | $\hat{s}(k-1)$ |
| 1 1 1 | -1.5375(1.5) -1.5375(1.5) | (a_1, a_1) | 1 |
| 1 1 -1 | -1.5375(1.5) -0.3875(-0.5) | (a_1, a_2) | 1 |
| -1 1 1 | 0.3875(0.5) -1.5375(1.5) | (a_3, a_1) | 1 |
| -1 1 -1 | 0.3875(0.5) -0.3875(-0.5) | (a_3, a_2) | 1 |
| 1 -1 1 | -0.3875(-0.5) 0.3875(0.5) | (a_2, a_3) | -1 |
| 1 -1 -1 | -0.3875(-0.5) 1.5375(-1.5) | (a_2, a_4) | -1 |
| -1 -1 1 | 1.5375(-1.5) 0.3875(0.5) | (a_4, a_3) | -1 |
| -1 -1 -1 | 1.5375(-1.5) 1.5375(-1.5) | (a_4, a_4) | -1 |

IV. FITNESS FUNCTION FOR OPTIMAL CHANNEL STATES

In order to find the optimal channel states, the use of the Bayesian likelihood (BL) [20] is considered. Since the Bayesian decision variable is a probability density function (pdf) variable, similar to the conventional likelihood, the BL can be defined by (9).

$$BL = \prod_{k=0}^{L-1} \max(f_B^{+1}(k), f_B^{-1}(k)) \quad (9)$$

where $f_B^{+1}(k) = \sum_{i=1}^{n_i+1} \exp(-\|y(k) - y_i^{+1}\|^2 / 2\sigma_e^2)$, $f_B^{-1}(k) = \sum_{i=1}^{n_i-1} \exp(-\|y(k) - y_i^{-1}\|^2 / 2\sigma_e^2)$ and L is the length of the received sequences. By evaluating the Bayesian likelihood, the optimal dataset of desired channel states which always corresponds to the maximum Bayesian likelihood would be found [13]. For this reason, the BL has been widely used as a fitness function (FF) in the previously developed search algorithms based on

GA or Fuzzy Clustering [13]-[16], and it is utilized as a fitness function for our modification of CFCM as well. Being more specific, the fitness function is taken as the logarithm of the BL , that is

$$FF = \sum_{k=0}^{L-1} \log(\max(f_B^{+1}(k), f_B^{-1}(k))) \quad (10)$$

Because of the characteristics of FF illustrated in [13] and [16], it cannot be easily solved by conventional gradient-based methods. On the other hand, the mathematical relation between FF and channel states cannot be formulated without the knowledge of channel structure [13]. Furthermore, it is too complex to be formulated even if the channel structure is known. Therefore in this paper, to search the optimal channel states which produce the maximum FF , under the presence of heavy AWGN, a modification of CFCM with Gaussian weighted partition matrix is developed.

V. MODIFICATION OF CFCM WITH GAUSSIAN WEIGHT

The conditional fuzzy clustering method was reported by W. Pedrycz in [17] and successfully applied to channel equalization problem [16][18]. In [17], the conditioning aspect of the clustering mechanism is introduced by taking into consideration the conditioning variable assuming values, f_1, f_2, \dots, f_k , on the corresponding patterns. This conditioning aspect, which describes a level of involvement of incoming input pattern in the constructed clusters, influences the clustering mechanism and improves the estimation accuracy of an unknown channel states for blind channel equalization [16]. Using the conditioning variables, f_1, f_2, \dots, f_k , makes it possible to apply the different weights to each of received patterns, which depend on their distances to the constructed clusters. To be more specific, the closer the received pattern to the clusters, the higher weight is attached and consequently more influential it becomes in the clustering process. For example, if $f_i = 0$, the i^{th} received pattern is regarded as meaningless in the clustering procedure and the calculations of the resulting prototypes are not affected by this element. Subsequently, the calculations of the partition matrix U in fuzzy clustering procedure do not take this into consideration. On the other hand, the pattern for which $f_i = 1$ contributes to the clustering process to the highest extent. This can be accomplished by the partition matrix U in CFCM derived as follows

$$U_{ik}^{(m+1)} = \frac{f_k}{\sum_{i=1}^{n_s} \left(\frac{\|y(k) - y_i^{(m)}\|}{\|y(k) - y_i^{(m)}\|} \right)^2} \quad (11)$$

$$y_i^{(m+1)} = \frac{\sum_{k=0}^{L-1} (U_{ik}^{(m+1)})^2 y(k)}{\sum_{k=0}^{L-1} (U_{ik}^{(m+1)})^2} \quad (12)$$

where $y_i^{(m+1)}$ is the i^{th} estimated center set at the $(m+1)^{th}$ iteration and $i=1,2,3,\dots,8$ for the channels in Table 1 because of $n_s=8$ (total number of desired channel states). By the same

way, the range of i for the channel in Table 2 is 1 to 16. Here in (11), the conditional constraint f_k should contain the distance information of each of received patterns, and it has a high value if the corresponding pattern is closely located at the estimated center. The CFCM in [16] utilizes each component of BL for the received patterns shown in (9) as the conditional constraint f_k after normalization, because it contains the distance information for each of received patterns. For an example, if a received pattern is located near the optimal desired channel states, y_i^{+1} or y_i^{-1} , this pattern produces a higher value of $f_B^{+1}(k)$ or $f_B^{-1}(k)$ in (9) and consequently it becomes more influential in the clustering process by (11). In other words, the closer the received pattern to the optimal channel states, the higher conditional constraint is applied. Because of the use of these conditioning variables, the performance of CFCM is relatively superior to those of the existing GA and FCM based approaches in terms of speed and accuracy. More details of the CFCM clustering algorithm for blind channel equalizations are described in [16]. The conditional constraint f_k of the CFCM represents the Gaussian probability value of each of received patterns because it depends on the BL in (9). However, the partition matrix U in (11) is still updated based on Euclidean distance measure. Because of AWGN, the received vector, $y(k)$, is scattered with a conditional Gaussian probability density centered at each of the desired channel states. Therefore, for the calculation of partition matrix U during the clustering procedure, the Gaussian probability of each of received patterns should be involved instead of the Euclidean distance measuring. The Gaussian weighted partition matrix U_G , where Euclidean distance is replaced with Gaussian probability, is described by (13) and a new center set y_i is sequentially derived by (14).

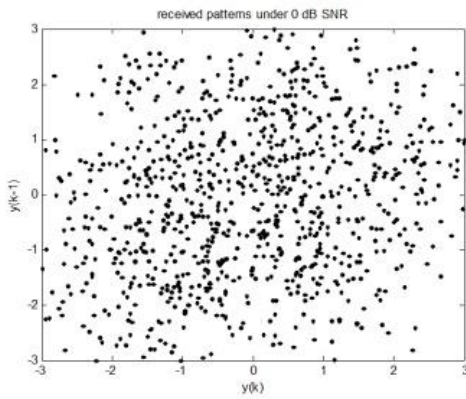
$$U_{ik}^{(m+1)} = \frac{f_k \cdot \exp(-\|y(k) - y_i^{(m)}\|^2 / 2\sigma_e^2)}{\sum_{i=1}^{n_s} \exp(-\|y(k) - y_i^{(m)}\|^2 / 2\sigma_e^2)} \quad (13)$$

$$y_i^{(m+1)} = \sum_{k=0}^{L-1} U_{ik}^{(m+1)} y(k) \quad (14)$$

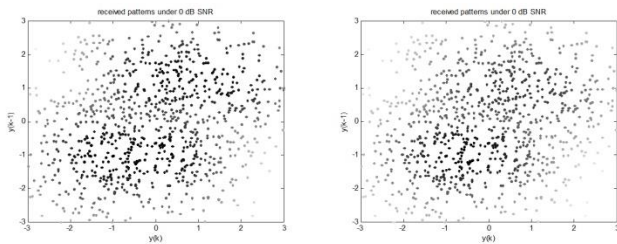
The effectiveness of the proposed Gaussian weighted partition matrix U_G under a heavy noise environment is demonstrated in Fig. 3. It shows the values of the conditional constraint f_k after 10 epochs of clustering procedure with the partition matrix U in (11) and U_G in (13) for the nonlinear channel in Table 1. The optimal centers of this channel are illustrated in Fig. 2(a). The received patterns under 0dB SNR are shown in Fig. 3(a), and the value of conditional constraint for each of those patterns is displayed by gray-colors (from 0(white) to 1(black)) in Fig. 3(b). For both cases, the noise-corrupted patterns, which are scattered and located far away from the optimal centers, have relatively very low constraint values (close to "0" indicated by bright color in Fig. 3(b)). On the other hand, the received patterns located near the optimal channel states are more weighted by the conditional constraint f_k (close to "1", black color in Fig. 3(b)) and generate higher contributions to the clustering procedure. However, the

received patterns with the high values of f_k in Fig. 3(c)-(e)_left are more widely spread than the patterns in Fig. 3(c)-(e)_right. In other words, it is observed that, in Fig. 3(c)-(e)_right, the high constraint values are assigned only for the received patterns which are more densely located near the optimal centers. It means that, by the clustering procedure with the partition matrix U_G , the closer located patterns near the optimal states have the relatively higher values of f_k than the values of f_k by clustering with U .

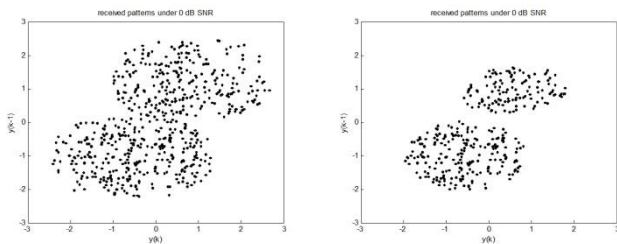
Therefore, in the proposed algorithm, the Gaussian weighted partition matrix U_G along with the conditional constraint f_k , shown in (13), is exploited instead of (11) and a new center set y_i is derived by (14). The resulting estimation accuracy is increased even with low SNRs and it is demonstrated in the next section.



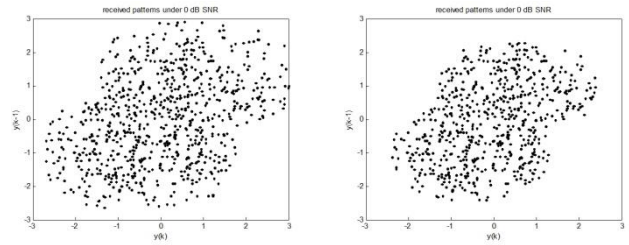
(a) received patterns under 0 dB SNR for the nonlinear channel in Table 1



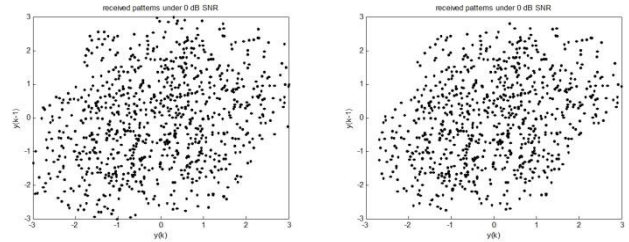
(b) received patterns displayed by f_k : 1(black) ↔ 0(white) (left: with U , right: with U_G)



(c) received patterns only for $f_k > 0.7$ (left: with U , right: with U_G)



(d) received patterns only for $f_k > 0.5$ (left: with U , right: with U_G)



(e) received patterns only for $f_k > 0.3$ (left: with U , right: with U_G)

Fig. 3. Received patterns under 0 dB SNR for the nonlinear channel in Table 1 and patterns displayed by their conditional constraint f_k (left: clustering with U , right: with U_G).

VI. SIMULATION RESULTS AND COMPARISONS

In order to demonstrate the performance of the proposed CFCM with U_G in search of the optimal channel states for blind channel equalization, the following simulations are carried out and compared. As mentioned in the introduction section, the performance of CFCM in [16] was superior to the previously developed GA based [13][14] and conventional FCM based [15] approaches in terms of speed and accuracy. Those algorithms also estimate the optimal channel states of an unknown channel to solve the blind equalization problem. Therefore the comparison for the effectiveness of the proposed method focuses on the CFCM in [16]. In the experiments, three channels including a linear model are evaluated. Channel 1 and 2 shown in Table 1 stand for each of nonlinear and linear models respectively, with the channel order $p=1$, and Channel 3 discussed in Section 3 concerns a nonlinear model with the channel order $p=2$ as presented in [16] and [21]. The first two channels were also often discussed in [13]-[16]. The detailed description of the channels is presented below.

Channel 1 (nonlinear): $H(z) = 0.5 + 1.0z^{-1}$,

$$D_1 = 1, D_2 = 0, D_3 = -0.9, D_4 = 0, \text{ and } d=1$$

Channel 2 (linear): $H(z) = 0.5 + 1.0z^{-1}$,

$$D_1 = 1, D_2 = 0, D_3 = 0, D_4 = 0, \text{ and } d=1$$

Channel 3 (nonlinear): $H(z) = 0.3482 + 0.8704z^{-1} + 0.3482z^{-2}$

$$D_1 = 1, D_2 = 0.2, D_3 = 0, D_4 = 0, \text{ and } d=1$$

In the experiments, 10 independent simulations for each of three channels with five different noise levels (SNR=0, 2.5, 5, 7.5, and 10dB) were performed with 1,000 randomly generated transmitted symbols ($L=1000$). Afterwards, the obtained results were averaged. The proposed CFCM with U_G and the ordinary CFCM with U have been implemented in a batch mode to facilitate a comparative analysis. In addition, both algorithms are evaluated with the use of the same parameters shown in Table 2, and these are fixed for all experiments.

The choice of the specific parameter values is not critical to the performance of both algorithms. For the evaluation purpose, the normalized root mean squared errors (NRMSE) is determined in the form

$$NRMSE = \frac{1}{\|a\|} \sqrt{\frac{1}{N} \sum_{i=1}^N \|a - \hat{a}_i\|^2} \quad (15)$$

where a is the data set of optimal channel output states, \hat{a}_i is the data set of estimated channel output states in the i^{th} simulation, and N is the total number of independent simulations ($N=10$).

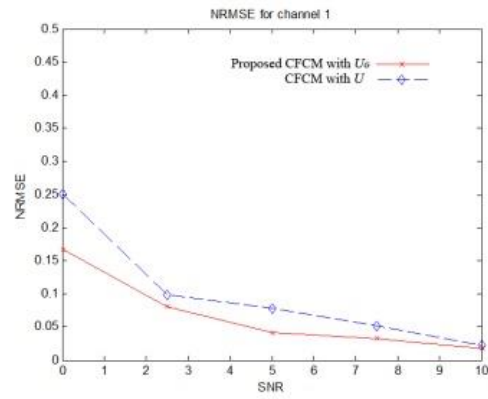
TABLE II. PARAMETERS USED IN SIMULATIONS

| | CFCM with U | Proposed CFCM with U_G |
|-----------------------------------|---------------|--------------------------|
| Maximum number of iteration | 100 | 100 |
| Threshold for FF variation | 10^{-3} | 10^{-3} |
| Exponent for partition matrix U | 2 | 1 |
| Random initial channel states | [-0.5 0.5] | [-0.5 0.5] |

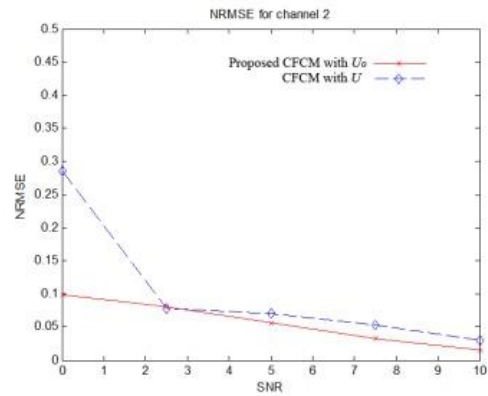
The values of NRMSEs after 10 independent simulations for each of three channels are averaged and illustrated in Fig. 4. The proposed CFCM with U_G comes with lower NRMSE for all three channels, and the performance differences are more severe in higher noise levels. As mentioned in the last part of Section 5 with Fig. 3, the clustering procedure with the Gaussian weighted partition matrix U_G in the proposed modification makes it possible that, the closer located patterns near the optimal states have relatively higher values of f_k than by clustering with U in the ordinary CFCM. Consequently, those patterns with higher f_k are more influential in the clustering process.

This effectiveness of Gaussian weighted partition matrix U_G is more critical in case of lower SNR because the type of corrupted noise in the channel is AWGN. Therefore the proposed CFCM with U_G is highly effective to find the optimal channel states when the received patterns are heavily corrupted by AWGN.

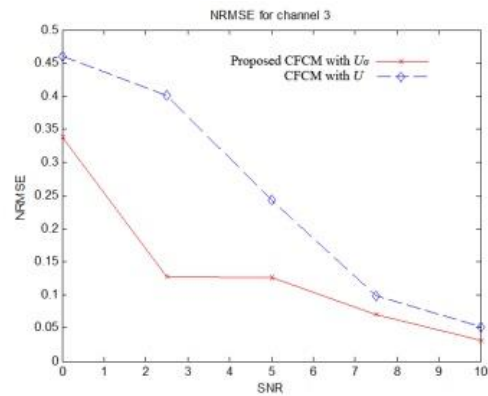
A sample of 1,000 received symbols under 0dB SNR for Channel 3 and its desired channel states constructed from the estimated channel output states by the proposed and the ordinary CFCM are illustrated in Fig. 5.



(a) for channel 1

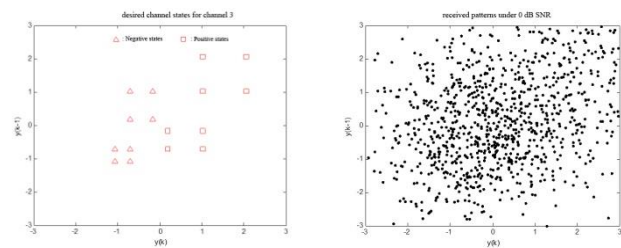


(b) for channel 2

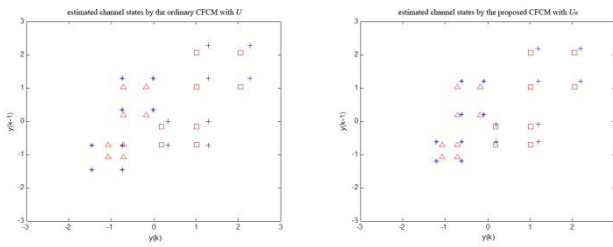


(c) for channel 3

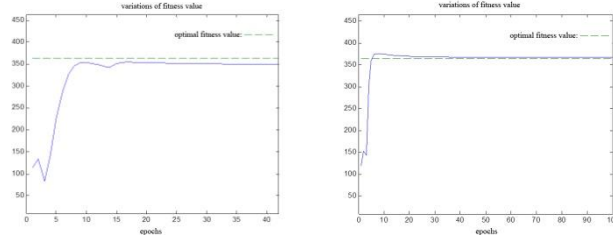
Fig. 4. NRMSEs by the proposed CFCM with U_G and by the ordinary CFCM with U .



(a) optimal channel states for channel 3 (b) received patterns under 0 dB SNR



(c) estimated states by CFCM with U (d) estimated states by CFCM with U_G



(e) fitness variations by CFCM with U (f) fitness variations by CFCM with U_G

Fig. 5. A sample of received symbols under 0dB SNR for Channel 3 and its sixteen desired channel states estimated by the proposed CFCM with U_G and the ordinary CFCM with U .

Because of the use of U_G , the proposed CFCM produces more accurate channel states from the noise-corrupted received patterns as shown in Fig. 5(d), and its fitness value by (10) during the clustering procedure approaches the optimal fitness value more closely as in Fig. 5(f). In addition, the relative search time (RST) of the proposed CFCM with U_G is evaluated. It is calculated by (16) and included in Table 4.

$$RST = \frac{\text{search time of CFCM with } U_G - \text{search time of CFCM with } U}{\text{search time of CFCM with } U} \quad (16)$$

As shown in Table 3, the values of RST for all three channels with different noise levels are almost zero, which means the search time difference between two algorithms is not significant where the proposed CFCM provides much better performance in terms of NRMSE. Additionally, some of RST for Channel 3, especially with low SNRs, are negative (faster search time for the proposed CFCM). It is caused by using the U_G in the clustering procedure, which reduces the number of convergence epochs in heavy noise circumstances.

TABLE III. RELATIVE SEARCH TIME (RST) FOR ALL THREE CHANNELS WITH DIFFERENT NOISE LEVELS

| Channel | SNR | RST |
|-----------|--------|---------|
| Channel 1 | 0.0 dB | 0.5588 |
| | 2.5 dB | 0 |
| | 5.0 dB | 0.0690 |
| | 7.5 dB | -0.1429 |
| | 10 dB | -0.1923 |
| Channel 2 | 0.0 dB | 0.4653 |
| | 2.5 dB | 0.2 |
| | 5.0 dB | 0 |
| | 7.5 dB | -0.1724 |
| | 10 dB | -0.1154 |
| Channel 3 | 0.0 dB | 0.2700 |
| | 2.5 dB | -0.4061 |
| | 5.0 dB | -0.2829 |
| | 7.5 dB | -0.4291 |
| | 10 dB | -0.2098 |

Finally, the bit error rates (BER) with the optimal and the estimated channel states are evaluated by using the Bayesian equalizer and they are summarized in Table 4. Even though the BER with the estimated channel states realized by the proposed CFCM with U_G is close enough to the one with the optimal channel states for all three channels, its performance does not dominate in terms of BER as much as it does in terms of NRMSE. Furthermore, especially for low SNRs, the BERs even with the optimal states are also relatively high. It is caused by the fact that the decision function of Bayesian equalizer shown in (7) is easily affected by heavy noise (high noise variance σ_e^2) even though the desired channel states can be estimated with high accuracy by using the proposed algorithm. For this reason, several nonlinear equalization techniques such as fuzzy or neural network implementations of Bayesian equalizer should be considered to improve the BER in next study.

TABLE IV. AVERAGED BER(%) (NO. OF ERRORS/NO. OF TRANSMITTED SYMBOLS).

| Channel | SNR | with optimal states | Ordinary CFCM with U | Proposed CFCM with U_G |
|-----------|--------|---------------------|------------------------|--------------------------|
| Channel 1 | 0.0 dB | 19.8 | 21.6 | 20.1 |
| | 2.5 dB | 15.3 | 15.4 | 15.4 |
| | 5.0 dB | 10.7 | 10.6 | 10.6 |
| | 7.5 dB | 6.69 | 6.81 | 6.80 |
| | 10 dB | 2.77 | 2.79 | 2.75 |
| Channel 2 | 0.0 dB | 19.3 | 21.6 | 19.2 |
| | 2.5 dB | 13.6 | 13.7 | 13.7 |
| | 5.0 dB | 8.95 | 9.17 | 9.08 |
| | 7.5 dB | 4.52 | 4.57 | 4.57 |
| | 10 dB | 1.79 | 1.76 | 1.76 |
| Channel 3 | 0.0 dB | 22.1 | 23.0 | 22.5 |
| | 2.5 dB | 16.1 | 16.9 | 16.5 |
| | 5.0 dB | 11.7 | 12.6 | 11.8 |
| | 7.5 dB | 7.94 | 8.22 | 7.88 |
| | 10 dB | 4.89 | 5.28 | 4.97 |

VII. CONCLUSIONS

The determination of an unknown channel states only from received patterns is critical in blind linear/nonlinear channel equalization problems. In this paper, for the estimation of desired channel states of an unknown digital channel under severe noise-corrupted communication environments, a modification of CFCM with Gaussian weighted partition matrix is presented and successfully evaluated with both of linear and nonlinear channels. Especially even when the received symbols are significantly corrupted by a heavy AWGN, it can estimate the channel output states with relatively high accuracy and substantial speed. Therefore, in the presence of heavy AWGN, the Bayesian equalizer with the proposed CFCM can be a possible solution for blind channel equalization. In future works, the evaluation of this method with higher order channels is included. In addition, as mentioned at the end of the last section, a further study on the implementation methods of Bayesian equalizer should be included to improve the BER under the presence of severe noise.

REFERENCES

[1] J. G. Proakis, Digital Communications, Fourth edition, McGraw-Hill, New York, 2001.

- [2] H. Gazzah and K. A. Meraim, "Blind ZF equalization with controlled delay robust to order over estimation", *Signal Processing*, vol.83, pp.1505-1518, 2003.
- [3] Yun Ye and Saman S. Abeysekera, "Efficient blind estimation and equalization of non-minimum phase communication channels via the use of a zero forcing equalizer", *Signal Processing*, vol. 86, pp.1019-1034, 2006.
- [4] J.R. Barry, E. A. Lee and D. G. Messerschmitt, *Digital Communication*, 3rd ed. Norwell, MA: Kluwer, 2004.
- [5] D. Erdogmus, D. Rende, J.C. Principe and T.F. Wong, "Nonlinear channel equalization using multilayer perceptrons with information theoretic criterion", *Proc. of IEEE workshop Neural Networks and Signal Processing*, pp. 443-451, MA, U.S.A., 2001.
- [6] N. Xie and H. Leung, "Blind equalization using a predictive radial basis function neural network," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 709-720, 2005.
- [7] M. Mimura and T. Furukawa, "A recurrent RBF network for non-linear channel," *Proc. of 2001 IEEE ICASSP*, vol. 2, pp.1297-1300, UT., U.S.A., 2001..
- [8] G. A. Barreto and L. G. M. Souza, "Adaptive filtering with the self-organizing map: A performance comparison," *Neural Networks*, vol. 19, no. 6, pp. 785-798, 2006.
- [9] S. Han, "Blind Equalization of Linear/Nonlinear Channels by SOM", *International Journal of Informaation Technology and Network Application*, vol. 2, no. 3, pp. 19-27, 2012.
- [10] A.K. Pradhan, S.K. Meher and A. Routray, "Communication channel equalization using wavelet network," *Digital Signal Processing*, vol. 16, no. 4, pp. 445-452, July, 2006.
- [11] B. Mitchinson and R. F. Harrison, "Digital communications channel equalization using the kernel adaline," *IEEE Transactions on Communications*, vol. 50, no. 4, pp. 571-576, 2002.
- [12] I. Santamaria, C. Pantaleon, L. Vielva and J. Ibanez, "Blind Equalization of Constant Modulus Signals Using Support Vector Machines", *IEEE Trans. Signal Processing*, vol. 52, pp.1773-1782, 2004.
- [13] H. Lin and K. Yamashita, "Hybrid simplex genetic algorithm for blind equalization using RBF networks", *Mathematics and Computers in Simulation*, vol. 59, pp.293-304, 2002.
- [14] S. Han, W. Pedrycz and C. Han, "Nonlinear Channel Blind Equalization Using Hybrid Genetic Algorithm with Simulated Annealing", *Mathematical and Computer Modeling*, vol. 41, pp.697-709, 2005.
- [15] S. Han, I. Lee and W. Pedrycz, "Modified fuzzy c-means and Bayesian equalizer for nonlinear blind channel", *Applied Soft Computing*, vol. 9, pp.1090-1096, 2009.
- [16] S. Han, S. Park and W. Pedrycz "Conditional fuzzy clustering for blind channel equalization", *Applied Soft Computing*, vol. 11, pp.2777-2786, 2011.
- [17] W. Pedrycz, "Conditional Fuzzy Clustering in the Design of Radial Basis Function Neural Networks", *IEEE Trans. Neural Networks*, vol. 9, pp.601-612, 1998.
- [18] K. Yoon, K. Kwak and S. Kim, "Nonlinear channel equalization using fuzzy clustering adaptive neuro-fuzzy filter", *Journal of Korea Electronics Engineers Society*, vol. 38, pp.35-41, 2001.
- [19] S. Chen, B. Mulgrew and S. McLaughlin, "Adaptive Bayesian equalizer with decision feedback", *IEEE Trans. Signal Processing*, vol. 41, pp.2918-2927, 1993.
- [20] H. Lin and K. Yamashita, "Blind equalization using parallel Bayesian decision feedback equalizer", *Mathematics and Computers in Simulation*, vol. 56, pp.247-257, 2001.
- [21] S.K. Patra and B. Mulgrew, "Fuzzy techniques for adaptive nonlinear equalization", *Signal Processing*, vol. 80, pp.985-1000, 2000.

An Object-Oriented Smartphone Application for Structural Finite Element Analysis

B.J. Mac Donald

Faculty of Engineering and Computing
Dublin City University, Dublin 9, Ireland

Abstract—Smartphones are becoming increasingly ubiquitous both in general society and the workplace. Recent increases in mobile processing power have shown the current generation of smartphones has equivalent processing power to a supercomputer from the early 1990s. Many industries have abandoned desktop computing and are now entirely reliant on mobile devices. Given these facts it is logical that smartphones are considered as the next platform for finite element analysis (FEA). This paper presents an architecture for a smartphone FEA application using object-oriented programming. A MVC design pattern is adopted and a demonstration FEA application for the Android smartphone platform is presented.

Keywords—Objected-oriented programming; Finite Element Method; Java; Android

I. INTRODUCTION

Since the introduction of smartphones in 2007 they have had a profound effect on lifestyles by significantly changing the way that people live, work and learn. Smartphones have become the dominant mobile device for communication information and entertainment. In many cases smartphones (and associated tablets) have become the dominant computing platform in many industries. Smith [1] demonstrates that in excess of 46% of American adults own a smartphone and the rate of ownership rises to in excess of 60% when college graduates and high income households (in excess of \$75,000) are considered. When considering these statistics, it is reasonable to assume that the majority of engineers, scientists and analysts will own, or have access to, a smartphone (or related tablet).

Many smartphone users are unaware of the computing power available in their devices and/or the potential of the smartphone as a platform for finite element analysis. Fig. 1 shows a comparison of computing power (in mega-flops) for different processors. The leftmost line (a) links the processing power of three supercomputers (Cray C1, Cray C90 and Cray Jaguar). The centre line (b) shows the processing power of desktop PC processors over time (Intel 386, Intel Pentium and Intel Core i7). The final line (c) illustrates the increase in computing power of mobile processors commonly used in Android smartphones and tablets. It is clear from fig. 1 that comparing a current high end mobile processor (e.g. Nvidia Tegra 4 which is built on ARM technology) with desktop and supercomputer processors, shows that current mobile processor capability is on par with desktop processors from circa 2008 and supercomputer processors from the early 1990's. Rajovic et al [2] discusses the development of mobile processor power

in comparison to supercomputers and suggests that multicore clusters of mobile processors may actually represent the future of high powered computing.

Given that fig. 1 shows that a current mobile device is approximately equivalent in computing power to an early 1990's supercomputer or a late 2000's desktop and, considering the pioneering finite element analyses work done on these machines at the time, it is reasonable to consider current smartphones as capable of performing useful finite element analyses.

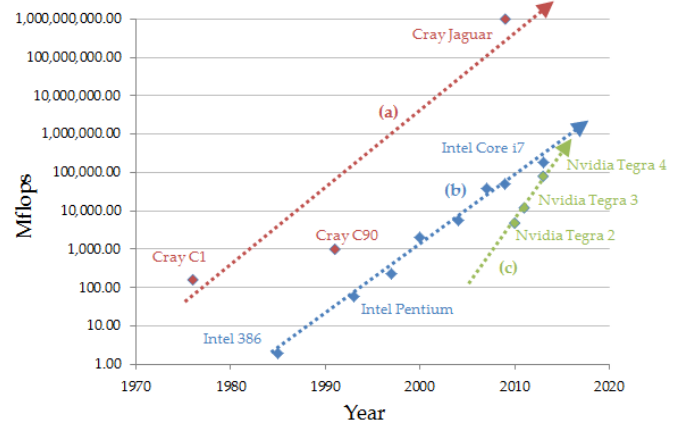


Fig. 1. Development in Computing Power (Mflops) since 1970. Trend lines show (a) supercomputers, (b) desktop PC's and (c) mobile processors.

There are currently two major operating systems available for smartphones: iOS (Apple Inc.) and Android (Google Inc.). Both of these platforms are based on objected-oriented programming languages: objective-C in the case of iOS and Java in the case of Android. Hence, any finite element code written for smartphones must be object-oriented.

Zimmermann et al. [3] described the governing principles for object-oriented finite element programming, before describing an implementation using SmallTalk [4] and C++ [5]. A number of authors [6-9] have described object-oriented implementations of the finite element method using C++.

Following the popularisation of Java in the late 1990's a number of researchers began to explore the possibilities of writing FEA codes using Java. Many researchers, however, were reluctant to engage with Java as it had a reputation for slow performance in comparison to more established non-object-oriented languages. In order to investigate this Nikishkov [10] compared the performance of a Java FEA code

with a similar code written in C. It was found that with the use of proper coding and tuning it is possible to obtain similar performance from the Java and C finite element codes. In a subsequent presentation, Nikishkov [11] described the design of an object-oriented Java finite element code for the 2D and 3D analysis of elastic and elasto-plastic structural solids. The code was developed using Java 1.5 and utilised the Java3D API to allow for visualisation of the results. A user interface was not developed and model specification was handled via an input text file which was read using a scanner.

This paper describes an object-oriented smartphone application written in Android, which is effectively a subset of Java. Android was chosen as it is an open source platform which runs on many devices including smartphones, tablets, netbooks and smart televisions. Graphical user interface (GUI) design on Android is relatively straightforward as the Android API contains a multitude of classes that can easily be subclassed to allow for complex displays and user inputs.

II. DESIGN OF THE FE APPLICATION

In order to simplify the discussion that follows we will initially consider a very simple finite element application that only solves 2D truss problems. The code outlined here may easily have additional classes defined which will allow the analysis of different structural problems using different types of finite element. The requirements for the application are shown in table I.

TABLE I. REQUIREMENTS FOR A SIMPLE SMARTPHONE FE APPLICATION

| No | Description |
|----|---|
| 1 | Function without error on the majority of Android devices |
| 2 | Use the device touchscreen to allow user input |
| 3 | Allow for FEA of 2D Truss problems |
| 4 | Allow the user to define nodes by their coordinates |
| 5 | Allow the user to define linear trusses by linking two nodes |
| 6 | Allow the user to define individual element properties |
| 7 | Allow the user to place a constraint on any node in either the x or y direction |
| 8 | Allow the user to place a force on any node in either the x or y direction |
| 9 | Allow the user to easily edit the model definition by changing properties |
| 10 | Easily and efficiently solve the finite element problem and present the results |
| 11 | Allow for sharing of the results via email, social media, etc. |

A Model-View-Controller (MVC) software architecture pattern was used to design the application. Fig. 2 shows an overview of the MVC pattern where we attempt to separate the representation of information from the interaction that the user has with this information. The *model* part of the pattern typically consists of data, logic and functions and, in this case, we can readily identify that our finite element classes belong here. We will designate a *model* package to contain the classes which describe the finite element model. The *view* part of the MVC pattern is used to output some representation of data to the user such as an image on a screen or a text listing etc. The *controller* part of the pattern takes input from the user and uses this input to send messages to the model or view. It is clear from fig. 2 that the user effectively interacts with the *view* part of the MVC pattern.

The view is also responsible for receiving user input and passing it to the controller. On a smartphone this is quite easy to grasp as the touchscreen on an Android device is used to both display the app and receive touch gestures. The controller receives user input from the view and acts accordingly. In most cases the controller will update the model state however it is also possible that the controller will just change the view without changing the model, for example, if a cosmetic change to the interface was requested by the user. The model stores data in its properties, implements application methods and implements the application logic. The model changes its state based on instructions from the controller. When the model changes its state it informs the view which updates accordingly.

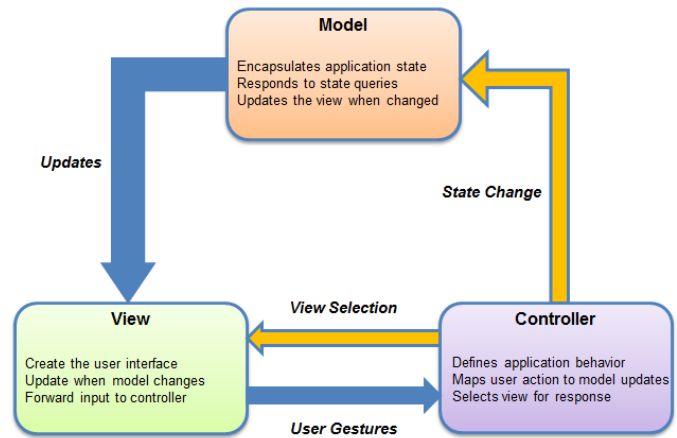


Fig. 2. Overview of the MVC Pattern

All Android applications must have a class designated as a “Main Activity” which is the entry point into the application – much in the same way as a class with a `main()` method is for a standard Java application. In this case we have named this class `TrussActivity` and this class must extend (i.e. be a subclass of) the `android.app.Activity` class. An Activity is an application component that provides a screen with which users can interact. By subclassing `android.app.Activity` our `TrussActivity` class will gain access to all the features of the Android API and be capable of displaying information on the device screen and receiving user input via touch gestures etc.

Android and Java classes are typically organised into Packages which contain classes that have a similar function or theme as discussed above. For illustration purposes we assume the package name: `com.example.simpletruss`. The `TrussActivity` class will be placed in this package making its full name: `com.example.simpletruss.TrussActivity`. Another package is used to hold the classes that may be used to define a finite element model. These classes are Java classes and are not specific to Android and hence may be reused for any Java application. In this case, a package named “model”: `com.example.simpletruss.model` is used to hold the finite element classes. Fig. 3 shows a basic schematic of the structure of the Android FEA app: illustrating the packages used and the classes which these packages contain.

The `TrussActivity` class will take user input and create objects from the classes contained in the `com.example.simpletruss.model` package and will call methods from these classes in order to build and solve the finite element model.

The classes within the `com.example.simpletruss.model` package are largely self-explanatory. The `Node` class is used to create `Node` objects and contains helper methods associated with the manipulation of `Node` objects. The `Truss2D` class is a subclass of the `LineElement` class which in turn is a subclass of the `Element` class. These classes are used to create `Element` objects. The `Assembly` class is used to create an assembly of finite elements and contains methods to create a global stiffness matrix, global force vector and a global nodal displacement vector. The `TrussSolver` class contains methods that can take these assembled global matrices and use them to obtain a solution to a finite element problem. The `TrussPost` class contains methods that can further process the obtained solution to obtain derived results such as element stress and strain. The `FeConstants` class contains a list of symbolic constants that may be used by all other classes within the package.

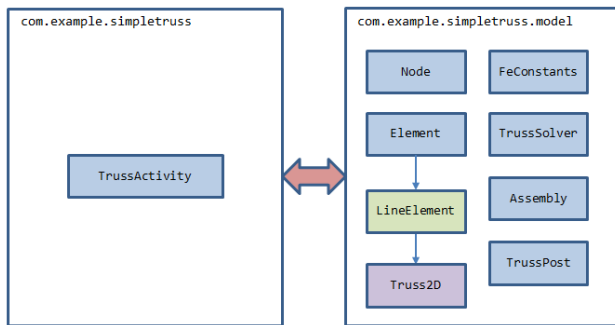


Fig. 3. Schematic of the Packages and Classes in the Android Application

By resolving the MVC pattern shown in fig. 1 with the schematic shown in fig. 2 it is clear that the `com.example.simpletruss.model` package will function exactly as the model is described in the MVC pattern. The `TrussActivity` class provides a method of linking into the Android API by subclassing `android.app.Activity`. Each Activity must implement the `onCreate()` method inherited from the superclass. In its simplest form the `onCreate()` method will be:

```
@Override
protected void onCreate(Bundle savedInstanceState) {
    super.onCreate(savedInstanceState);
    setContentView(R.layout.main_layout);
}
```

The final line in the above code snippet calls the superclass method `setContentView()` to set the View that will be shown to the user when the application is started. This layout is usually specified in an XML layout file named `main_layout.XML`. This XML layout file may be edited to display the buttons, text fields, checkboxes, images, etc. that make up the applications GUI.

Effectively, the `onCreate()` method links us into the Android API via `android.app.Activity` and provides us with a View using `setContentView()` via `android.view.View`.

The controller part of the MVC pattern will consist of the other methods contained in `TrussActivity` which are not inherited from the superclass. These are methods which are custom written for the FE application. These methods are summarised in table II.

TABLE II. CONTROLLER METHODS IN TRUSSACTIVITY

| Method | Description |
|---------------------------------|---|
| <code>addNode()</code> | Creates a Node object using user input from a dialog box |
| <code>deleteNode()</code> | Deletes a Node object from the database using a dialog box |
| <code>addElement()</code> | Creates an Element object using a dialog box |
| <code>deleteElement()</code> | Deletes an Element object from the database |
| <code>addConstraint()</code> | Sets a constraint on a Node object using a dialog box |
| <code>deleteConstraint()</code> | Modifies a constraint on a Node object |
| <code>addForce()</code> | Sets a force on a Node object using a dialog box for user input |
| <code>deleteForce()</code> | Modifies a force on a Node object |
| <code>calculate()</code> | Uses the database of Node and Element objects to create an assembly of finite elements, solves the global problem and then creates a new View to display the results, simultaneously saves the results to a text file for sharing |

Each of the methods described in Table II performs two basic functions: instructing the view what view to provide (add a node dialog, delete a force dialog, results screen etc.) and processing user input from this view and using it to change the state of the model (add a new node object, change the force on a node object, etc.).

So, in summary, the `com.example.simpletruss.model` package contains the *Model*, the `onCreate()` method in `TrussActivity` class and its associated XML files contain the *View* and the other methods in `TrussActivity` class define the *controller*. This is illustrated in fig. 4

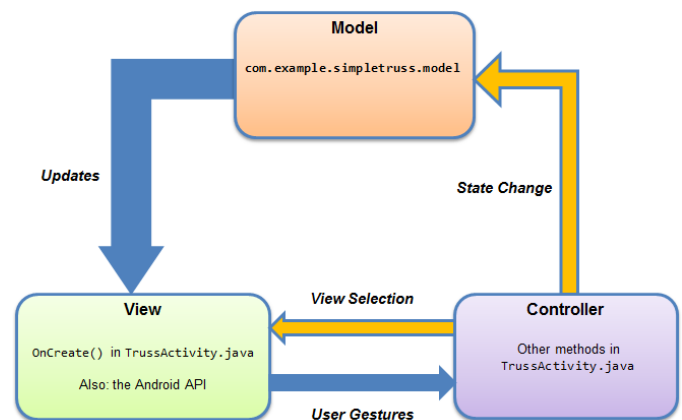


Fig. 4. A MVC Implementation for the FE Smartphone Application

III. MODEL

The subsections below describe the classes in the model package which is responsible for building the finite element model and solving the global problem.

A. Node

The Node class is common to all finite element types and will be unchanged regardless of the element type used. A description of the Node class is shown in table III.

TABLE III. DESCRIPTION OF THE NODE CLASS

| Node |
|--|
| - id: int - coord: double [] - force: double [] - bc: double [] |
| + Node(int id, double x, double y, double z) + getCoords(): double[] + setCoords(double x, double y, double z): void + getID(): int + setID(int id): void + getBc(): double[] + setBc(double x_bc, double y_bc, double z_bc): void + getForce(): double[] + setForce(double x_force, y_force, z_force): void + isLoaded(): boolean + isX_const(): boolean + isY_const(): boolean + isZ_const(): boolean + drawNode(Canvas canvas, Paint paint, float scale, int alpha): void + drawBC(Canvas canvas, float scale): void + drawLoads(Canvas canvas, float scale): void |

Each node object has an id, array of global coordinates, array of applied forces and an array of boundary conditions. Each of these arrays has three members of double precision numbers: for the x, y and z directions.

The Node constructor creates a node object using its id and its x, y and z coordinates. Nodes may be defined in 2D space by setting z equal to zero. Public getter and setter methods are provided in order to allow for reporting and modification of a nodes properties. A number of helper methods are provided to quickly determine if a node has an applied load or boundary condition. These methods return a Boolean value and are generally used to aid in the graphical display of loads and boundary conditions. Finally, helper methods are provided which allow for display of the node and its associated applied forces and boundary conditions in the applications GUI.

B. Element Classes

The Element class is an abstract class for all finite element types. A description of the Element class is shown in table IV. Each element object must have an id, a list of nodes that define the element and an elastic modulus. Several abstract methods (shown in italics) are provided which must be implemented by any subclasses: these methods provide for reporting of element properties and assembly of the elements stiffness matrix and strain displacement matrix.

TABLE IV. DESCRIPTION OF THE ELEMENT ABSTRACT CLASS

| Element |
|---|
| # elemId: int # nodelist: int [] # elasticModulus: double |
| + <i>getElemId(): int</i> + <i>getNodeList(): int[]</i> + <i>getElasticModulus(): double</i> + <i>stiffnessMatrix(): double [][]</i> + <i>strainDispMatrix(): double [][]</i> |

The LineElement class is a subclass of the Element class and, as such must implement its abstract methods. Table V shows a description of the LineElement class.

TABLE V. DESCRIPTION OF THE LINEELEMENT CLASS

| LineElement |
|--|
| # node1 : Node # node2 : Node # crossSectionArea : double |
| + LineElement(int eld, Node n1, Node n2) + setElemId(int eld): void + getElemId(): int + setNode1(Node n1): void + getNode1(): Node + setNode2(Node n2): void + getNode2(): Node + setElasticModulus(double eMod): void + getElasticModulus(): double + setCrossSectionArea(double csa): void + getCrossSectionArea(): double + elementLength(): double + lcos(): double + mcos(): double |

Each LineElement object is defined by two Node objects and its cross sectional area. The constructor creates LineElement objects using this data. Several getter and setter methods are provided to allow for reporting and modification of element properties. Finally, three helper methods are provided which calculate the element length and its direction cosines, l and m .

The Truss2D class is a subclass of both LineElement and Element (via the class hierarchy). A 2D truss is obviously a line element and so inherits all the properties and methods of its superclass. The Truss2D class is primarily concerned with implementing methods specific to 2D truss finite elements. Table V shows a description of the Truss2D class. Since most of the methods required for a 2D truss have already been implemented in the superclass's, the Truss2D class is relatively short. It essentially consists of a constructor which simply calls the constructor of the superclass and two methods which calculate the element stiffness matrix and strain displacement matrix.

TABLE VI. DESCRIPTION OF THE TRUSS2D CLASS

| |
|---|
| Truss2D |
| - stiffnessMatrix : double [4][4] - strainDispMatrix : double [1][4] |
| + Truss2D(int eld, Node n1, Node n2) + stiffnessMatrix() : double [4][4] + strainDispMatrix() : double [1][4] |

C. Assembly Class

The Assembly class essentially consists of five methods which assemble the global problem equations. The global stiffness matrix is assembled from the individual element stiffness matrices and placed into a 2D array of double precision numbers. Similarly the global force vector and global displacement vector are assembled by interrogating each element to find its constituent Node objects and their relevant force and boundary condition data. Two further methods are used to assemble global data which will be useful during post-processing of results. An ArrayList of element strain-displacement matrices and an ArrayList of element elastic-modulii are produced by calling these methods. The Assembly class is written in a non-element specific manner so that it may be used with any element type, not just the truss elements being considered here.

D. TrussSolverClass

The TrussSolver class contains one method named calculateDisplacements() which returns the solved nodal displacement vector to the calling method or class. A direct equation solver performs solution of the system equation using symmetric LDU decomposition of the matrix.

E. TrussPost Class

The TrussPost class contains a number of methods for post-processing the results from a truss analysis. The strains() method is used to return an array of doubles which effectively gives the strain in each element in the finite element model. Similarly, method stress() provides an array listing the axial stress in each element in the finite element model. Finally, method reactionForces() is used to return an array listing the reaction forces at each node in the finite element model.

F. A Note on the Model Classes

Clearly, there are several possibilities available for class construction and interaction when using an objected-oriented approach. The above description attempts to take the four principles of object-oriented design (Encapsulation, inheritance, polymorphism and abstraction) into account at all times. It could be argued that the Assembly, Solver and Post-Processor classes could be either combined into one class, or, are not really classes at all and their methods should be combined into other classes (e.g. one of the element classes). Alternatively, these methods could be placed in a class which contains only a list of static methods and thus does not require instantiation in order to call the methods. Both of these strategies, however, would remove the flexibility of the software and make it more difficult to add additional element types to the finite element application.

IV. VIEW

As described in section II, each screen in Android is represented using an XML layout file. It is also possible to create the layout dynamically during program execution but, in most cases, it is preferable to define an XML layout in advance. Fig. 5 shows the main screen used for building the finite element model in the completed smartphone application.

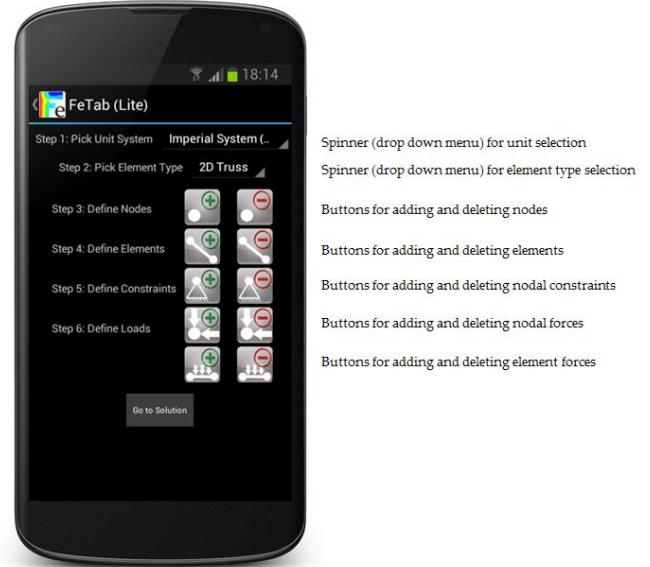


Fig. 5. Graphical User Interface (GUI) for the Smartphone Application Pre-Processor

The screen layout is divided into a number of steps that the user is required to complete in order to successfully build a finite element model. The layout was constructed in this manner in order to avoid user confusion and also, as one of the aims of the application was for it to be used as an educational tool to teach FEA to new users. In the first step a drop down menu (known as a “Spinner” in Android) is used to capture the user’s preference in terms of unit system. The selected unit system is used to prompt the user for input quantities during the model generation and also during the display of results. The user is offered three choices: no units (which is the default), the SI system (Kg-m-sec) or the Imperial system (lb-ft-sec). Step 2 requires the user to pick an element type: currently there are three options available: 1DTruss, 2DTruss and Beam. The class system for a 2D truss analysis was discussed in section III. A 1D truss can be easily formed by simply setting the appropriate coordinates and DOF to zero. A beam element was implemented by adding additional classes to the structure discussed in section III and, for the sake of clarity, will not be discussed here. Step 3 requires the user to specify nodal coordinates. Touching on either the add node or delete node button opens a dialog box which allows the user to define the nodal coordinates. Similarly the add element, delete elements, add constraint, delete constraint, add nodal force and delete nodal force buttons all open appropriate dialog boxes for the user to interact with. The two lower buttons allow for the application of distributed loads if a beam element type has been selected – if a truss element is selected then these buttons

will display a warning. Fig. 6 shows examples of the “Add Node” and “Add Element” dialogs.

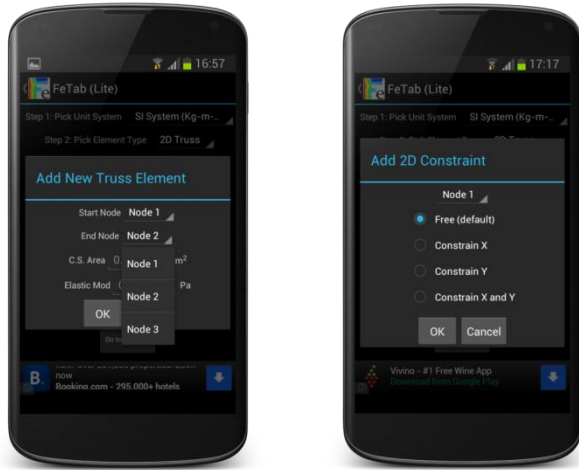


Fig. 6. Dialog Boxes are Used to Capture User Input

During the XML definition of the buttons shown in fig. 5 a method name in TrussActivity is required in order to link the button to that method. For example the “Add a Node” button definition contains a reference to the addNode() method in TrussActivity. When the button is touched/clicked then the relevant method is called and the object reference of the View calling the method is passed as a parameter to the method.

The full suite of Android’s user interface was utilised to capture input from the user: including spinners, checkboxes, radio-buttons, textboxes etc. Touching the application icon at the top of the screen slides a menu out from the left hand side which allows the user to navigate through the application, as shown in fig. 7.

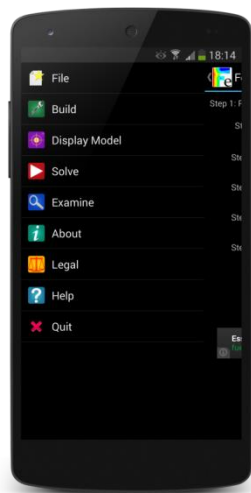


Fig. 7. Smartphone FE Application Navigation Menu

Touching the “File” button allows the user to load or save a model and clear the database. The “Build” button brings up the pre-processor screen shown in fig. 5. The “Display Model” button is used to show a graphical representation of the model, as shown in fig. 8. The “Examine” button is used for post-processing the results from the finite element model. The other buttons in fig. 7 are largely self-explanatory.

Each of the screens described above are created using XML layout files which specify the relative position of the various UI elements. These layouts are displayed by the corresponding Android activity class when required. In some cases, such as with dialogs, the display is created dynamically using only Java code without the need for a XML layout to be defined in advance. This is achieved using one of the many “builder” classes provided with the Android API. The graphical display of the model is also created dynamically by filling an empty frame layout with a Canvas object when the user requests the model be displayed.

Fig. 8 shows a typical graphical display from a 2D Truss problem. In this case three nodes and two elements have been used. Node numbers are displayed near the associated nodes. Constraint and load symbols are placed on relevant nodes, using the helper methods described in section III. A facility for zooming in/out and an option to fit the finite element model to the screen are provided in the lower right corner of the GUI.



Fig. 8. Graphical Display of a 2D Truss Problem. Note display of constraint symbols on left hand side and load arrow symbols on the right.

V. CONTROLLER

The TrussActivity class is the main activity for a truss analysis. As mentioned above, the first task of TrussActivity is to call the onCreate() method from its superclass. This method is called when the application is started and is responsible for providing the View for the Activity by linking to the appropriate XML layout file.

The main task of TrussActivity is to act as controller in the MVC pattern and to take user input in order to use the Model classes to construct a finite element model. Table VII shows a description of the TrussActivity class, focusing on the methods dealing with control. A number of EditText object references are initially described as private class variables. EditText's are editable text boxes that are used to obtain user input. In this case they are required to capture nodal coordinates, element properties, etc. Two ArrayList objects are defined which effectively act as the finite element model database. The nodes ArrayList holds a list of currently defined Node objects and the elements ArrayList holds a list of currently defined LineElement objects. ArrayLists are effectively mutable arrays and so allow for the addition and subtraction of objects from the list as required. Two integer variables are defined in order to conveniently hold the number of currently defined nodes and elements.

TABLE VII. DESCRIPTION OF THE TRUSSACTIVITY CLASS

| TrussActivity |
|---|
| <ul style="list-style-type: none">- xNode : EditText- yNode : EditText- startNode : EditText- endNode : EditText- area : EditText- elasticModulus : EditText- deleteNode : EditText- deleteElement : EditText- xConst : double- yConst : double- nodes : ArrayList<Node>- elements : ArrayList<LineElement>- numNodes : int- numElements : int |
| <ul style="list-style-type: none"># onCreate(Bundle savedInstanceState) : void+ addNode(View v) : void+ delNode(View v) : void+ addElement(View v) : void+ deleteElement(View v) : void+ addConstraint(View v) : void+ deleteConstraint(View v) : void+ addForce(View v) : void+ deleteForce(View v) : void+ calculate (View v) : void+ printTrussResults(View v) : void |

The addNode() method is triggered by the user touching the "Add a Node" button on the main screen (fig. 5). An object reference to the View that requested the method to be called is passed in as the parameter v. This reference is required as it tells the addNode() method how/where to update the View if required.

Each of the methods shown below the addNode() method in table VII follow a standard procedure so the addNode() method will be used to illustrate this procedure. The addNode() method begins by creating a dialog box in the current View in order to obtain user input. The method then sets up a listener

to listen for either the cancel or OK buttons in the dialog box to be touched by the user. If the cancel button is touched then the dialog is simply dismissed and control is returned to the calling method. If the OK button is touched then data entered by the user is checked for viability. If the data is not viable then a message is displayed to the user explaining why this is the case. If the data is viable then a new Node object is created using the object constructor in the Node class. This Node object is then added to the nodes ArrayList and the numNodes variable is incremented by 1 before returning control to the calling method.

Some of the other methods require more checks before displaying a dialog requesting user input. The addElement() method, for example, first checks that at least two Node objects have been defined before allowing the user to proceed. In each case where a problem is encountered an explanatory message is presented to the user.

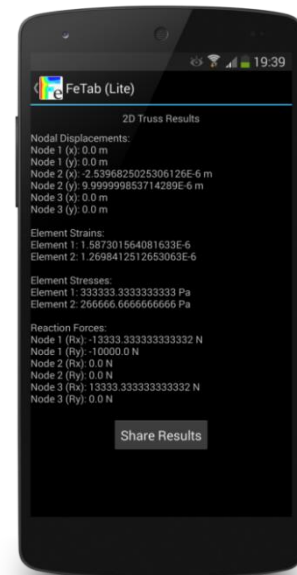


Fig. 9. Typical Results Display

The calculate() method begins solution of the finite element model. Before attempting to form the global assembly a number of checks are carried out to ensure the model is ready for solution: at least one element is defined, at least on DOF is constrained, at least one nodal force has been specified, etc. In each case an appropriate message is displayed to the user if a problem is encountered. If no problems are found then assembly of the global system of equations proceeds as described in section III. The assembled problem is then solved using the TrussSolver class which returns an Array containing the solved global displacement Vector. A quick check is performed to ensure that the returned array is not empty (indicating a failed solution). If this is the case then a message regarding the mathematical un-stability of the finite element model, together with some advice on how to fix the model is presented to the user. If the global displacement vector is valid

then the `printTrussResults()` method is called and results are automatically post-processed and displayed to the user. Fig. 9 shows a typical display of results from a simple 2D truss analysis.

VI. CONCLUSION

The architecture of a demonstration finite element analysis application for an Android smartphone has been presented. The application has been designed according to object-oriented principles using a MVC design pattern. Smartphone user interfaces provide exciting opportunities to revolutionise the generation and analysis of finite element models. In this case the objective was to produce a functioning finite element application which could also be used as an educational tool to teach new users basic FEA principles. The Android platform makes it relatively easy to design an intuitive and educational user interface. The architecture provided here can easily be expanded to include more complex elements and analysis capabilities. The demonstration application is available for free download [12].

REFERENCES

- [1] A. Smith, 46% of American Adults are now Smartphone Owners, Pew Internet, 2012 (<http://pewinternet.org/Reports/2012/Smartphone-Update-2012.aspx>)
- [2] N. Rajovic, P. Carpenter, I. Gelado, N. Puzovic and A. Ramirez, Are Mobile Processors Ready for HPC?, edaWorkshop13, Dresden, Germany, May 14-16, 2013.
- [3] T. Zimmermann, Y. Dubois-Pélerin and P. Bomme, Object-oriented Finite Element Programming: I. Governing Principles, Computer Methods in Applied Mechanics and Engineering, 1992, 98, No. 2, pp. 291-303.
- [4] T. Zimmermann, Y. Dubois-Pélerin and P. Bomme, Object-oriented Finite Element Programming: II. A Prototype Program in Smalltalk, Computer Methods in Applied Mechanics and Engineering, 1992, 98, No.3, pp. 361-397.
- [5] T. Zimmermann, Y. Dubois-Pélerin and P. Bomme, Object-oriented Finite Element Programming: II. An Efficient Implementation in C++, Computer Methods in Applied Mechanics and Engineering, 1993, 108, No.1-2, pp. 165-183.
- [6] P. Donescu & Tod. A Laursen, A Generalized Object-Oriented Approach to Solving Ordinary and Partial Differential Equations Using Finite Elements, Finite Elements in Analysis and Design, 1996, 22, pp. 93-107
- [7] J. Besson & R. Foerch, Large Scale Object-oriented Finite Element Code Design, Computer Methods in Applied Mechanics and Engineering, 1997, 142, pp. 165-187.
- [8] G.C. Archer, G. Fenves & C. Thewalt, A New Object-oriented Finite Element Analysis Program Architecture, Computers and Structures, 1999, 70, pp. 63-75
- [9] B. Patzák & Z. Bittnar, Design of Object-oriented Finite Element Code, Advances in Engineering Software, 2001, 32, 759-767
- [10] G.P.Nikishkov, Yu.G.Nikishkov and V.V.Savchenko, Comparison of C And Java Performance In Finite Element Computations, Computers and Structures, 2003, 81, pp. 2401-2408
- [11] G.P.Nikishkov, Object oriented design of a finite element code in Java. Computer Modeling in Engineering and Sciences, 2006, 11, No. 2, pp. 81-90
- [12] <https://play.google.com/store/apps/details?id=ie.jion.fetab>

New Approach for Image Fusion Based on Curvelet Approach

Gehad Mohamed Taher

Computer Science
Faculty of Computer and Informatics
Suez Canal University
Ismailia, Egypt

Mohamed ElSayed Wahed

Computer Science
Faculty of Computer and Informatics
Suez Canal University
Ismailia, Egypt

Ghada EL Taweal

Computer Science
Faculty of Computer and Informatics
Suez Canal University
Ismailia, Egypt

Abstract—Most of the image fusion work has been limited to monochrome images. Algorithms which utilize human colour perception are attracting the image fusion community with great interest. It is mainly due to the reason that the use of colour greatly expands the amount of information to be conveyed in an image. Since, the human visual system is very much sensitive to colours; research was undertaken in mapping three individual monochrome multispectral images to the respective channels of an RGB image to produce a false colour fused image. Producing a fused colour output image which maintains the original chromaticity of the input visual image is highly tricky.

The focus of this paper is developing a new approach to fuse a color image (visual image) and a corresponding grayscale one (Infrared image – IR) using the curvelet approach using different fusion rules in new fields. The fused image obtained by the proposed approach maintain the high resolution of the colored image (visual image), incorporate any hidden object given by the IR sensor as an example, or complements the two input images and keep the natural color of the visual image.

Keywords—Image fusion; visual colored image; monochrome images

I. INTRODUCTION

Image fusion is a process of combining complementary information from multiple sensor images to generate a single image that contains a more accurate description of the scene than any of the individual images. As for example while MMW (millimeter wave) sensors have many advantages, the low cost IR makes the study of fusing visual and IR images of great interest.

In our work, we are using fusion to help human or computer to detect the hidden objects using IR and visual sensors. Most of the image fusion work has been limited to monochrome images [1]. However, based on biological research results, the human visual system is very sensitive to colors. Waxman, Aguilar [4-6]. Al use a neural network to fuse a low- light visible image and IR image to generate a three channel false color image used for night operations. In addition, Aguilar [7] has extended their work to fuse multi-modality volumetric imagery.

In this paper, we introduce a new approach to fuse a color visual image with a corresponding grayscale IR image or any other sensor; using the proposed approach the fused image will

maintain the high resolution and the natural color of the visual image.

The paper is organized as follows, section 2 describes the proposed image fusion approach, and section 3 presents the experimental results which demonstrate the feasibility of the proposed fusion approach, section 4 conclusion.

II. CURVELET TRANSFORM

The curvelet transform is a very young signal analysing method with good potential. It is recognized as a milestone on image processing and other applications [2].

So why we use curvelet; actually, the time frequency analysis is decomposed a signal to several orthogonal bases. We can quantize the signal to the summation of different basis with different coefficient:

$$f = \sum_k a_k b_k \quad (1)$$

↑ ↑
Curvelet Coefficients basis. frame

Curvelet transform is more accurate to deal with the curve than wavelet transform the below Fig. 1 shows this.

Wavelet approach

Many wavelet coefficients are needed to account edges.

i.e. singularities along lines or curves needed to account edges.

Curvelet approach

Less coefficients are needed to account edges.

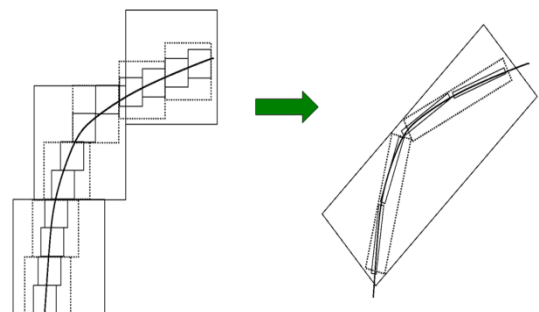
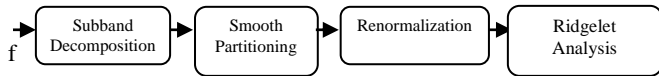


Fig. 1. Difference between curvelet and wavelet approach

The overview of the curvelet transform is shown below for four step:



A. Subband Decomposition

We define a bank of subband filter P_0 , ($\Delta s, s \geq 0$). The object f is filter into subbands:

$$f \mapsto (P_0 f, \Delta_1 f, \Delta_2 f, \dots)$$

This step divides the image into several resolution layers. Each layer contains details of different frequencies:

$P_0 \rightarrow$ Lowpass filter

$\Delta_1, \Delta_2, \rightarrow$ Band-pass (high-pass) filters

B. Smooth Partitioning

Let w be a smooth windowing function so by applying it to the decomposition we get

$$h_Q = w_Q \cdot \Delta_s f \quad (2)$$

C. Renormalization

In this stage of the procedure, each ‘square’ resulting in the previous stage is renormalized to unit scale:

$$g_Q = T_Q^{-1} h_Q \quad (3)$$

D. Ridgelet construction

The ridgelet construction divides the frequency domain to dyadic coronae Fig.2. In the angular direction, it samples the s -th corona at least $2s$ times. In the radial direction, it samples using local wavelets. [3]

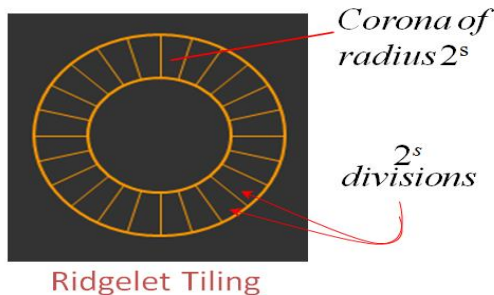


Fig. 2. Ridgelet Tiling and Fourier

Transform of the Curvelet Transform each normalized square is analyzed in the ridgelet system:

$$\alpha_{(Q,\lambda)} = \langle g_Q, \rho_\lambda \rangle$$

III. FUSION APPROACH FOR GRAY AND VISUAL IMAGES

The proposed image fusion approach is illustrated in Fig. 3.

This method is based on the color space transform RGB – HSI – RGB; The HSI is based on the RGB true color space. An RGB color image is given by an $M \times N \times 3$ array of color pixels, where each pixel is a triplet corresponding to the red, green and blue component of an RGB image at a specified location.

In HSI color space, the hue component represents the dominant color present in an image. The saturation component indicates the amount of purity. The intensity component gives the gray level values of the image.

The HSI color system is considerably closer than RGB system to human perception in describing the color sensations. Further, HSI color space allows the de-coupling of intensity component from the color carrying information in an image. Hence HSI color space is used for intermediate processing in an image fusion task.

A. Steps of the proposed approach:

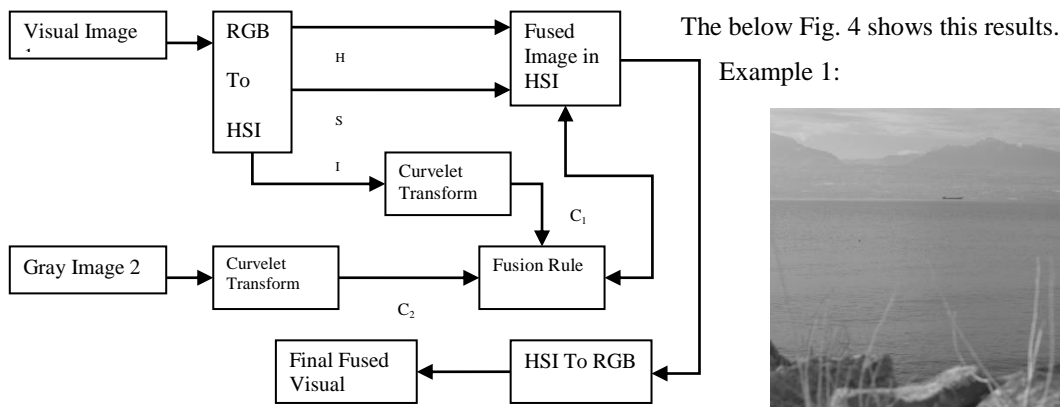
- 1) Input two images of the same scene one grayscale and one color image.
- 2) The RGB components of color image (Image 1) are converted to HSI components where

$$I = (R + G + B)/3$$

$$H = (B - R)/3(I - R), S = (1 - R)/I \text{ when } R = \text{Minimum}(R, G, B)$$

$$H = (R - G)/3(I - G), S = (1 - G)/I \text{ when } G = \text{Minimum}(R, G, B)$$

$$H = (G - B)/3(I - B), S = (1 - B)/I \text{ when } B = \text{Minimum}(R, G, B)$$
- 3) Curvelet transform is applied to intensity of image 1 and the other grayscale image 2 (IR image) respectively using the Wrapping Algorithm.
- 4) Three different fusion rules are applied to the coefficients that are PCA, wavelet, and mean fusion rules at each location of the input images to produce a single set of coefficients in the fused output.
- 5) Inverse curvelet transform is applied to get the final fused output.
- 6) Now hue, saturation components are added to the intensity image to get the final fused color image.
- 7) Finally, the HSI color space is converted to RGB format.



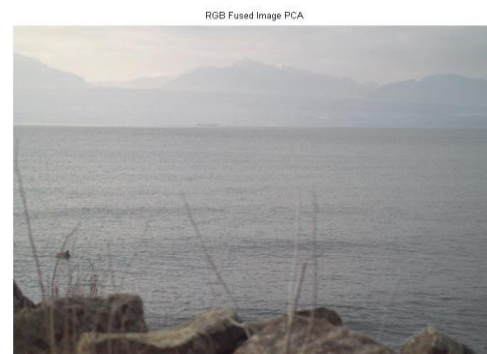
input image1 monochrome



Input image2 visual



RGB Fused Image mean



RGB Fused Image PCA

Fig. 3. Block diagram for the proposed approach

B. Fusion Rules

There are a variety of fusion rules that have been reported as valid image fusion processes. Some of the popular fusion techniques based on statistical analysis of the images that used in our test is mean, and Principle Component Analysis (PCA). Assuming that images are collected simultaneously with accurate registration, images can be fused element wise, taking the mean values. PCA is an orthogonal linear transformation technique that transforms the multidimensional data sets to lower dimensions for image analysis without much loss of information content. The new coordinate system obtained by PCA transformation is such that the greatest variance by any projection of the data lies in the first coordinate (principle component).the second greater variance on the second coordinate, so on. We use the popular wavelet based approach to find the decomposition coefficients for image fusion. The wavelet based method is available as the image fusion tool in the wavelet toolbox, which is used for fusing various registered images of the same size. The principle of image fusion using wavelets is to merge the wavelet decompositions of the two original images using fusion methods. [2]

IV. EXPERIMENTAL RESULTS

In order to illustrate the efficiency of the proposed image fusion approach, a dozen images were used in the experimental tests.

A. Data Set

By applying the experimental test for a dataset for more than 10 pair of images one is visual and the other is gray scale with the same size, as the resolution of the visual image is much higher than that of the gray scale, but the visual one doesn't convey all of the information in the gray one. So by fusing both we got a new one that is high resolution from the visual and all of the things appear from the gray one.

B. Qualitative performance comparison

By applying the test to a pair of 4 images we found that first the visual appearance of the resulted fusion image is much better than the two input image for the three different fusion rules mean, PCA, wavelet.



Fig. 4. Two input images of example1 and the result for applying mean, PCA and wavelet respectively

In this example the mean fusion rule is more accurate and have better resolution than the two other fusion rule.

Example 2:



input image1 monochrome



Input image2 visual



Fig. 5. Two input images of example2 and the result for applying mean, PCA and wavelet respectively

As a result for this fusion the mean fusion rule get all the details from the greyscale image but light the colours of the visual image, the other two fusion rule didn't get all the fine details from the greyscale one.

C. Quantitative performance comparison

Quantitative performance is the one which involve predefined quality indicators for measuring the spectral and the spatial similarities between the fused image and the original image. [3] Here we use quality measures like mean square error, peak signal to noise ratio, entropy and standard deviation. [9]

a) Root Mean square error

The RMSE between a reference image R and the fused image F is given by There are different approaches to construct reference image using input images. In our experiments, we used the following procedure to compute RMSE. First, RMSE value E1 is computed between source image A and fused image F.

$$E1 = \sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (R(i, j) - F(i, j))^2} \quad (4)$$

Smaller RMSE value indicates good fusion quality.

b) Peak signal to noise ration

The ratio between maximum possible power of the signal to the power of the corrupting noise that creates distortion of image. The peak signal to noise ratio can be represented as

$$PSNR = 10 \times \log_{10} \left(\frac{255 \times M \times N}{\sum \sum (x(i, j) - y(i, j))^2} \right) \quad (5)$$

Where x- fused image, y – perfect image, i – pixel row index, j – pixel column index, M, N – Number of rows and columns respectively.

Entropy:

The entropy of an image is a measure of information content .The estimate assumes a statistically independent source characterized by the relative frequency of occurrence of the elements in X, which is its histogram. For a better fused image, the entropy should have a larger value. A high value of entropy denotes more information content and vice versa. [8]

$$H(S) = -\sum P(X) \log P(X) \quad (6)$$

Standard deviation:

The standard deviation (SD) provides a way to determine regions which are clear and vague, it is the square root of variance, reflects the spread in the data. Thus, a high contrast image will have a larger variance, and a low contrast image will have a low variance. [8]

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (7)$$

Where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (8)$$



Fig. 6. Two input images one greyscale and the other visual image of the country scene

Applying these metrics to the first pair of images in Fig. 6 we get those results in table 1

TABLE I. RESULTS OF APPLYING RMSA, PSNR, ENTROPY AND SD TO IMAGES OF FIG. 6

| | RMSA | PSNR | Entropy | SD |
|---------------------|-------------|---------|---------|--------|
| PCA fusion rule | 1.4237e+004 | 6.5966 | 7.6036 | 0.2384 |
| Mean fusion rule | 664.5194 | 19.9057 | 7.2583 | 0.1685 |
| Wavelet fusion rule | 2.1859e+003 | 14.7346 | 7.0864 | 0.1304 |

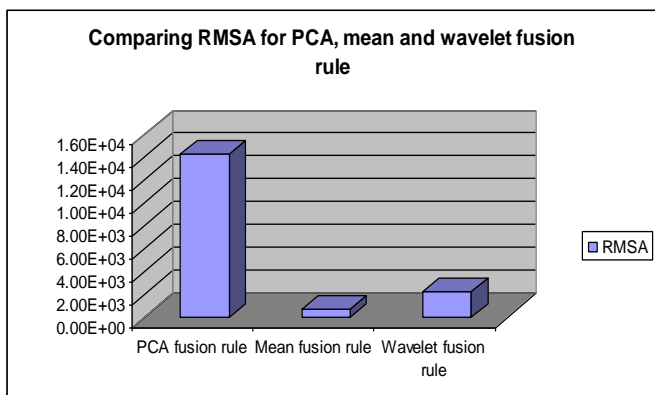


Fig. 7. RMSA for PCA, mean and wavelet fusion rule for table 1, mean fusion gives the low RMSA indicating the best fusion quality than the PCA and wavelet fusion rule

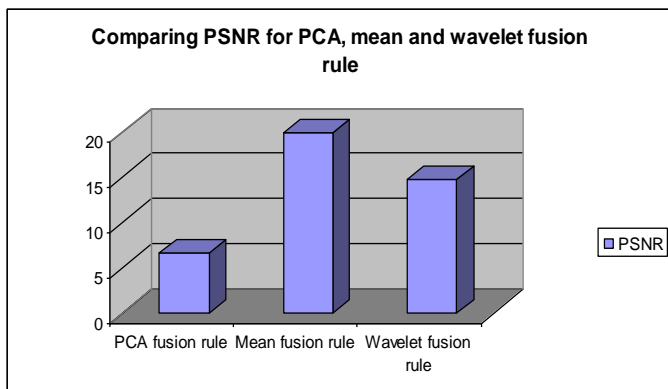


Fig. 8. PSNR for PCA, mean and wavelet fusion rule, mean fusion rules gives high PSNR indicating more contrast and less noise image than PCA and wavelet fusion

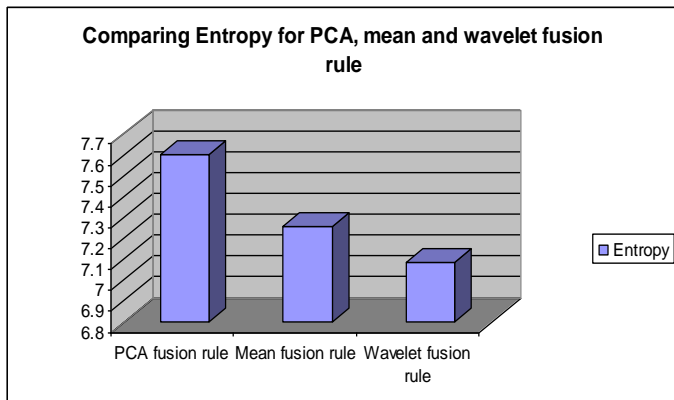


Fig. 9. Entropy for PCA, mean and wavelet fusion rule, PCA is with high entropy than mean and wavelet fusion rule.

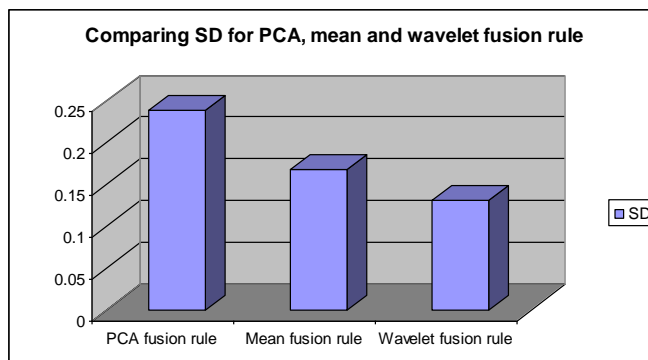


Fig. 10. SD for PCA, mean and wavelet fusion rule PCA is much contrast than mean and wavelet fusion rule

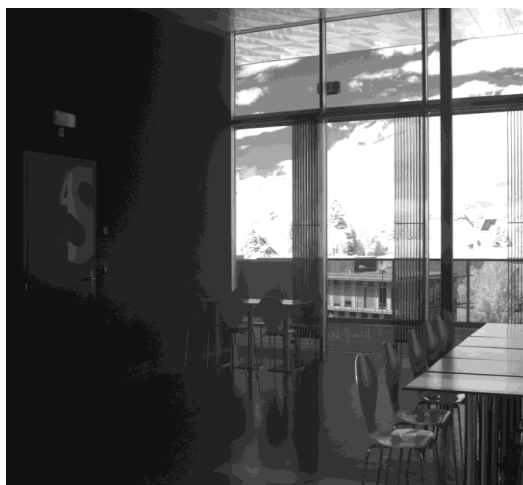


Fig. 11. Two input images for indoor scene one greyscale and the other visual

Applying these metrics to the second pair of images in Fig. 11 we get those results in table 2

TABLE II. RESULTS OF APPLYING RMSA, PSNR, ENTROPY AND SD TO IMAGES OF FIG. 11

| | RMSA | PSNR | Entropy | SD |
|----------------------------|--------------------|----------------|---------|--------|
| PCA fusion rule | 595.3910 | 20.3828 | 7.2164 | 0.3130 |
| Mean fusion rule | 1.6878e+003 | 15.8577 | 7.0609 | 0.3389 |
| Wavelet fusion rule | 533.4170 | 20.8601 | 7.2073 | 0.3120 |

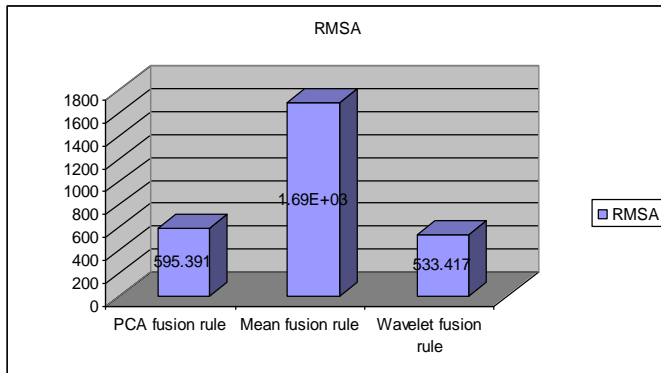


Fig. 12. RMSA for PCA, mean and wavelet fusion rule for table 2, wavelet fusion gives the low RMSA indicating the best fusion quality than the PCA and mean fusion rule

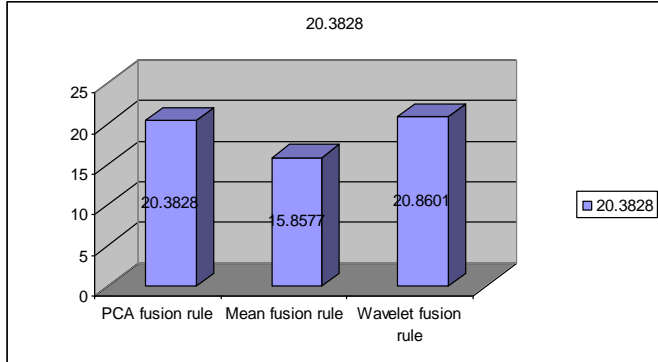


Fig. 13. PSNR for PCA, mean and wavelet fusion rule, wavelet fusion rules gives high PSNR indicating more contrast and less noise image than PCA and wavelet fusion

V. CONCLUSION

In this paper we proposed a new image fusion approach for combining a visual colored image and a corresponding grayscale one having some hidden objects or more accurate in some parts of the corresponding visual one; for enhancing the input images we use the curvelet transform approach and used different fusion rules and show a comparison for a three different fusion rule, as for a future work we can use different fields and different type of sensors and apply them to another fusion rules to provide a fused image that provides a detailed description of the details of the scene that appear in one image and not the other and more accurate and high resolution than the input ones.

REFERENCES

- [1] M. Wahed, Ghada EL Taweal, A. Fouad, Gehad M. Taher, Image fusion approach with noise reduction using genetic algorithm, International Journal of Advanced Computer Science and Applications, Vol. 4, No. 11, 2013
- [2] Karthik P. Ramesh, Shruti Gupta, and Erik P. Blasch, "Image Fusion Experiment for Information Content", Information Fusion, 10th International Conference on, Vol. 9, 2007.
- [3] Y. Zhang, Methods for image fusion quality assessment - a review, comparison and analysis, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXVII (B7) (2008) 1101-1110.
- [4] A. M. Waxman, M. Aguilar, R. A. Baxter, D.A. Fay, D. B. Ireland, J. P. Racamato, W. D. Ross, Opponent-color fusion of multi-sensor imagery: visible, IR and SAR, Proceedings of IRIS Passive Sensors, vol.1, pp. 43-61, 1998.
- [5] M. Aguilar, D. A. Fay, W. D. Ross, A. M. Waxman, D. B. Ireland, J. P. Racamato, Real-time fusion of low-light CCD and uncooled IR imagery for color night vision, Proceedings of the SPIE, vol.3364, pp. 124-35, 1998.
- [6] A. M. Waxman, M. Aguilar, D. A. Fray, D. B. Ireland, J. P. Racamato, Jr., W. D. Ross, J. E. Carrick, A. N. Gove, M. C. Seibert, E. D. Savoye, R. K. Reich, B. E. Burke, W. H. McGonagle, D. M. Craig, Solid-state color night vision: fusion of low-light visible and thermal infrared imagery, Lincoln Laboratory Journal, vol. 11, no. 1, pp. 41-60, 1998.
- [7] M. Aguilar, and J. R. New, Fusion of multi-modality volumetric medical imagery, ISIF 2002, pp. 1206-1212.
- [8] Manu V T and Philomina Simon, A Novel Statistical Fusion rule for image fusion and its comparison in non subsampled contourlet transform domain and wavelet domain: The International journal of multimedia & its Applications (IJMA) Vol.4, No.2, April 2012.
- [9] Y.Kiran Kumar Technical Specialist- Philips HealthCare, Comparison Of Fusion Techniques Applied To Preclinical Images: Fast Discrete Curvelet Transform Using Wrapping Technique & Wavelet Transform. Journal of Theoretical and Applied Information Technology

Automated Menu Recommendation System Based on Past Preferences

Daniel Simon Sanz¹, Ankur Agrawal¹

¹Department of Computer Science, Manhattan College, Riverdale, NY 10471

Abstract—Data mining plays an important role in ecommerce in today's world. Time is critical when it comes to shopping as options are unlimited and making a choice can be tedious. This study presents an application of data mining in the form of an Android application that can provide user with automated suggestion based on past preferences. The application helps a person to choose what food they might want to order in a specific restaurant. The application learns user behavior with each order - what they order in each kind of meal and what are the products that they select together. After gathering enough information, the application can suggest the user about the most selected dish in the recent past and since the application started to learn. Applications, such as these, can play a major role in helping make a decision based on past preferences, thereby reducing the user involvement in decision making.

Keywords—data mining; Apriori; Android; restaurant; recommendation system

I. INTRODUCTION

E-commerce has become an important aspect of our life today. The popularity of mobile devices has further helped towards pushing this trend upwards [1]. People are spending more time shopping online on their mobile devices today than they were in the past. At the same time, inventories of online stores are ever increasing giving shoppers more options to choose from. However, with the availability of so many different options, making a decision can be stressful. Data mining can play an important role in providing user with the right information at the right time, thus improving the shopping experience of the user [2].

Data mining is being used in several ecommerce sites such as Amazon and ebay. However, restaurants are still lagging behind in the application of data mining to improve their operations. Relationships and trends in user data can be identified and used to improve user experience by providing them with automated menu recommendations based on their past preferences.

This paper describes a study performed by applying data mining techniques and Android application development techniques in creating an application for a restaurant. The application is a digital menu that differentiates the kind of meal based on the time of the day and suggests the user the most common dishes for that kind of meal. The application uses Apriori algorithm [3] to analyze the information that the user enters and provides automated recommendations based on past

usage. As such, the application is constantly learning about the user and evolving with each usage.

Android [4] is a mobile operating system based on the Linux kernel. It has been designed for devices with a touchscreen such as smartphones and tablets. Android was developed by Android Inc. that was founded in 2003. It was bought by Google in 2005 [5]. Android is the world's most powerful and popular mobile platform with over a million new Android mobile devices activated every day [6].

The Apriori algorithm is a data-mining algorithm for frequent mining of item sets over transactional databases. The algorithm works by analyzing a dataset considering a minimum support threshold. The algorithm then identifies the individual items with a frequency greater or equal than the threshold, and creates datasets by combining all those items. The algorithm does the same with the new datasets, until there is any item in the set that has more frequency than the minimum support threshold.

II. METHODS

The application is programmed in Android, API level 19 (for Android 4.4.2 or higher), using eclipse [7] as the integrated development environment. SQLite3 [8] is used as the backend which is a simple way to save user data.

The MySQLHelper class is used to create and fill the database. The database has three different tables as shown in Figure 1. The "Dishes" table has the following attributes - id as the primary key, name of the dish, the description, the category and the price. The second table is called "Transactions" and contains the id as the primary key, the kind of meal and the date of the transaction. The "Transactions Dishes" table has an id and also holds the primary keys of the first two tables as foreign keys. This table is used to provide data to the Apriori algorithm which provides the data mining capabilities to the application.

Next, there are three classes working as content providers [9], one for each of the three tables in the database. Content providers manage access to a structured set of data. They encapsulate the data, and provide mechanisms for defining data security. They are also the standard interface that connect data in one process with code running in another process. In this case, these classes are used to provide an easy way to communicate and interact with the database such as the ability to insert a row in any of the database tables.

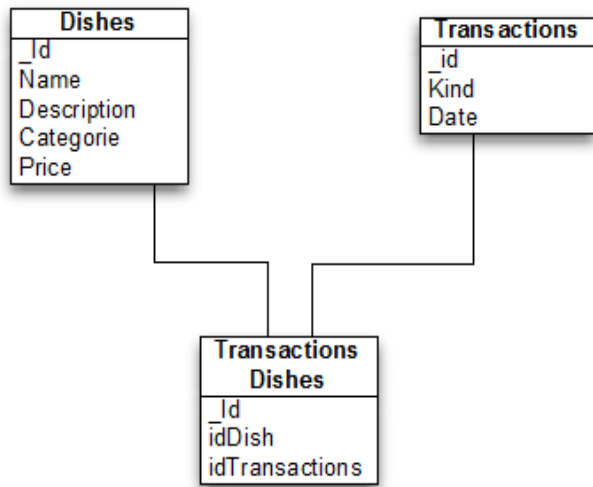


Fig. 1. Database structure

There are two activities [10] included with the application. An activity is a screen where the programmer establishes the GUI, in order to interact with the user. Normally, these activities are full screen but they can also appear in other forms such as floating windows, or embedded inside of another application.

The activities consist of spinners [11]. A spinner provides a quick way to select a value from a set. In the default state, a spinner shows its currently selected value. Touching the spinners display a dropdown menu with all other available values from which the user can select a new one.

The Apriori algorithm is used to identify the names of the dishes that the user usually purchases together. The minimum threshold for the algorithm is two. The algorithm keeps track of the user selection and after gathering enough information about user behavior, it is able to recommend a dish to the user based on their past preferences.

III. RESULTS

Figure 2 displays a screenshot of a cellphone with the application icon named “DeliChooser” in the last row. Clicking on the icon opens the application.

The application has two activities. Figure 3 displays the first activity of the application in its default state. The first activity consists of three spinners. The first spinner is used to determine the kind of meal such as dinner, breakfast, lunch, etc. This field is set automatically depending on the hour of the day. However, the user is also given the ability to change it for each transaction. The second spinner is used to determine the category of the dish, such as breakfast platters, breakfast wraps, desserts, hot signature sandwiches, etc. The third spinner is used to choose a dish such as chicken parmesan, blondie brownie, ghost rider, etc.



Fig. 2. Screenshot showing the application icon

The second and third spinners are related because every time the user changes the category using the second spinner the new list of dishes is automatically shown using the third spinner.

The application also has six text views following the three spinners. The first one is a static text view and its only function is to show the string “Selected:” whenever a dish is selected. The second and third text views show the most selected dish in the last month as well as historically depending on the kind of meal selected in the first spinner. The fourth and the fifth text views show characteristics of the dishes, and change every time the third spinner is modified. The fourth text view shows the price of the dish and the fifth text view shows the description of the dish. The sixth text view appears only when the user chooses a dish, and the application has enough information to display about the dishes that the user normally selects with the dish that was chosen.

Figure 4 displays all the six text views after the user has made a selection. The first text view is “Selected:”, the second text view is “Historically: Chicken Parmesan”, the third text view is “Last Month: Chicken Parmesan”, the fourth text view is “\$7.99”, the fifth text view is the description “Seasoned Italian sausage grilled multi colored bell peppers marinara sauce, mushrooms & onions, served on a hero bread” and the sixth text view is “You usually choose the following dish(es) with this one: - Classic Breakfast Wrap”. The sixth text view appears when the user clicks on the button “Select” which also saves the current transaction into the database.

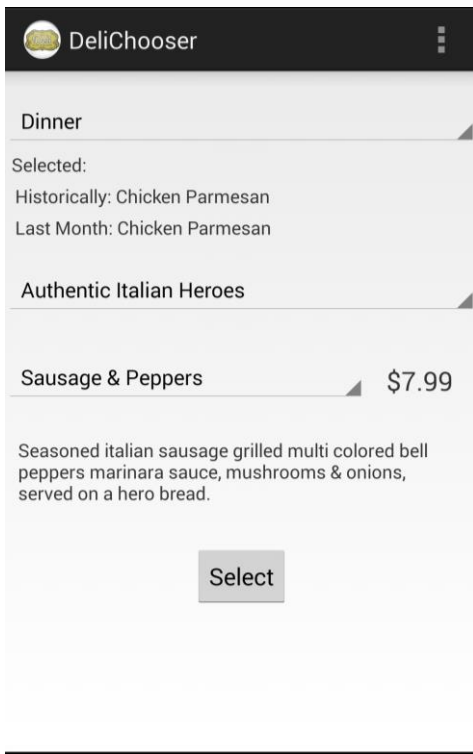


Fig. 3. Activity 1 (Default)

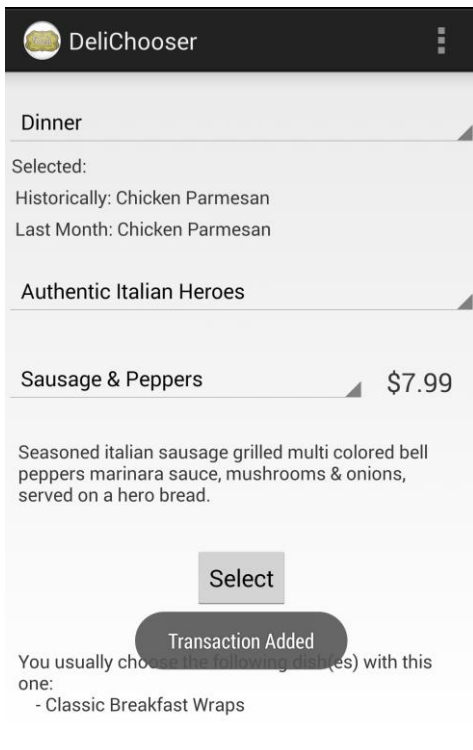


Fig. 4. Activity 1 (Transaction Added)

The second activity is used to display information about the restaurant as shown in Figure 5. It provides the name and logo, hours, address, and telephone number of the restaurant.

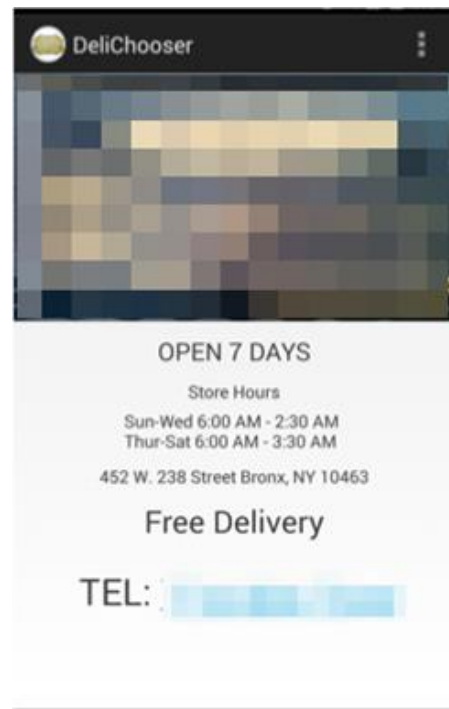


Fig. 5. Activity 2 (Default)

IV. DISCUSSION

With the ever growing product inventory in e-commerce sites, an average user has to browse through several pages before they decide to buy a product. This can be overwhelming and tedious. Recommendation systems help user by providing them with suitable matches as per their preferences, thereby, saving hours of their search time.

This paper presented one such recommendation system for a restaurant environment. The application keeps track of a user's past preferences in selecting a menu item and after gathering sufficient information, the system is able to suggest the user with menu items that the user might prefer. The application also takes into consideration the time of the day when making a recommendation. This removes the hassle of browsing through the menu to decide what to order thus saving user their valuable time.

The application was developed using Android operating system as it is the most popular operating system and used by majority of the mobile users. A lot of companies have their own Android application which makes it easier for the users to access their information faster than in any other ways, because they have all that they need in their cell phones which can be used on the go. Data mining techniques such as the Apriori algorithm was utilized to develop the recommendation system. The algorithm keeps track of the user behavioral data which is an indication of what items does the user normally purchase together and how frequently is that done by the user. By considering such factors and by learning over time, the algorithm is able to make smart recommendations to the user thus saving the user valuable time of having to browse through the menu every time.

Future work will involve adding and displaying nutritional information about the dishes, and also adding information about multiple restaurants. A study will also be performed to further improve the algorithm by adding other factors related to user purchase pattern such as the price of the item and the nutritional value of the item that the user normally purchases. A webpage will also be developed and connected to the application database which will allow the restaurant manager to keep track of the items that the customers normally purchase together. This will enable them to make relevant offers, add new dishes and change prices. Future work will also involve applying the recommendation algorithm to other fields such as a trip recommendation system.

V. CONCLUSION

Recommendation systems have become an important part of our digital life. These systems are being used by several e-commerce sites to improve user experience of online shoppers as they can be a valuable resource to the user. This paper presented a menu recommendation system for a restaurant. The application provided user with menu recommendation depending on the users past preferences and the time of the day. The Android operating system was used to develop the application and the Apriori algorithm was used as the basis of the recommendation system. Applications such as this can

contribute towards improving the shopping experience of the users by saving them both time and effort.

REFERENCES

- [1] L. Einav, J. Levin, I. Popov and N. Sundaresan, Growth, adoption and use of mobile e-commerce, *The American economic review*, 104(5), pp. 489-94, 2014.
- [2] Data mining in ecommerce, [retrived July 2014] <http://www.ias.ac.in/sadhana/Pdf2005AprJun/Pe1299.pdf>
- [3] R. Agrawal and R. Srikant, Fast algorithm for mining association rules, In Proc. 20th int. conf. very large databases, VLDB (Vol. 1215), pp. 487-499, September 1994.
- [4] Android, [retrived July 2014] <http://www.Android.com/>
- [5] Android Wikipedia Page, [retrived July 2014] [http://en.wikipedia.org/wiki/Android_\(operating_system\)](http://en.wikipedia.org/wiki/Android_(operating_system))
- [6] About Android, [retrived July 2014] <http://developer.android.com/about/index.html>
- [7] Eclipse, I.D.E., The Eclipse Foundation, 2007, [retrived July 2014] <https://www.eclipse.org/>
- [8] SQLite, [retrived July 2014] <https://www.sqlite.org/>
- [9] Android Content Provider, [retrived July 2014] <https://developer.Android.com/guide/topics/providers/content-providers.html>
- [10] Android Activity, [retrived July 2014] <https://developer.Android.com/reference/Android/app/Activity.html>
- [11] Android Spinner, [retrived July 2014] <https://developer.Android.com/guide/topics/ui/controls/spinner.html>

A Shape Based Image Search Technique

Aratrika Sarkar¹

¹Department of Computer Science and Engineering
Institute of engineering and Management,
West Bengal University of Technology,
Kolkata, India.

Pallabi Bhattacharjee¹

¹Department of Computer Science and Engineering
Institute of engineering and Management,
West Bengal University of Technology,
Kolkata, India

Abstract—This paper describes an interactive application we have developed based on shaped-based image retrieval technique. The key concepts described in the project are, i) matching of images based on contour matching; ii) matching of images based on edge matching; iii) matching of images based on pixel matching of colours. Further, the application facilitates the matching of images invariant of transformations like i) translation; ii) rotation; iii) scaling. The key factor of the system is, the system shows the percentage unmatched of the image uploaded with respect to the images already existing in the database graphically, whereas, the integrity of the system lies on the unique matching techniques used for optimum result. This increases the accuracy of the system. For example, when a user uploads an image say, an image of a mango leaf, then the application shows all mango leaves present in the database as well other leaves matching the colour and shape of the mango leaf uploaded.

Keywords—*shape-based image retrieval; contour matching; edge matching; pixel matching*

I. MOTIVATION

Humans can often recognize objects using shape information alone. This has proven to be a challenging task for computer vision systems. One of the main difficulties is, developing representations that can effectively capture important shape variations. We want to compare different objects such as leaves of a tree and to detect two different leaves. The computational complexity of these tasks and the recognition accuracy obtained are highly dependent on the choice of a shape representation and comparison among leaves of various shapes.

II. INTRODUCTION

In this 21st century where “searching” on the internet in a part and parcel of life, text and voice search are dominant whereas image search is still lagging behind. So we have tried to develop an application which shows optimum results on Image Searching.

Recognition [1] relies upon the existence of a set of predefined objects. Content-based image retrieval [2] is prevalent since 1992 for automatic retrieval of images from a database, based on the colors and shapes present. Since then, the term has been used to describe the process of retrieving desired images from a large collection on the basis of syntactical image features [3]. The techniques, tools, and algorithms that are used originate from fields like statistics, pattern recognition, signal processing, and computer vision [4].

The problem of recognition of objects represented in images is the problem of identifying homologous elements in shapes, which are usually defined by groups of points. Our approach focuses on finding the optimum matching of the images taking contour [5] as the key feature of the image. Contour matching [6] is an important issue and a difficult problem of image processing. The accuracy and the efficiency of the algorithms are the most two critical factors. Contour representation defines the boundary of an object. We must keep in mind that the object must be identified even if it undergoes some geometric transformations. We aim to find the output as images which match the input image in terms of maximum percentage matching. The user has options to find out results in terms of EDGE MATCHING [7], CONTOUR MATCHING, COLOUR MATCHING [8]. In section III we discuss the related work, followed by the Methodology presented in section IV. Section V presents some snapshots while Section VI shows the experimental results and finally in Section VII we conclude the discussions.

III. RELATED WORK

Many methodologies have been proposed to analyze plant leaves in an automated fashion. A large percentage of such works utilize shape recognition techniques to model and represent the contour shapes of leaves, however additionally, color and texture of leaves have also been taken into consideration to improve recognition accuracies. One of the earliest works employs geometrical parameters like area, perimeter, maximum length, maximum width, elongation to differentiate between four types of rice grains, with accuracies around 95% [9]. Use of statistical discriminant analysis along with color based clustering and neural networks have been used for classification of a flowered plant and a cactus plant. The authors use the Curvature Scale Space (CSS) technique [10] and k-NN classifiers [11] to classify chrysanthemum leaves. Both color and geometrical features have been reported to detect weeds in crop fields employing k-NN classifiers. The authors propose a hierarchical technique of representing leaf shapes first by their polygonal approximations and then introducing more and more local details in subsequent steps. Fuzzy logic [12] decision making has been utilized to detect weeds in an agricultural field. The authors propose a two-step approach of using a shape characterization function called centroid-contour distance curve [13] and the object eccentricity [14] for leaf image retrieval. The centroid-contour distance (CCD) curve and eccentricity along with an angle code histogram (ACH) [15] have been used for plant recognition.

IV. METHODOLOGY

Fig 1 shows the flowchart for the algorithm developed. There are several functions which the application can perform. It allows the user to input an image of his/her choice. Then the user has the facility of matching edge, contour which will take the image as input and perform several functions. For instance "Find edges" is a function, which will take an image as input,

and find out the matching images as output whose edges match the edge of the input image. Contour representation defines the boundary of an object. The object has to be identified even if it undergoes some geometric transformation and our application succeeds in finding out the matching of the input image even if it undergoes geometric transformations [16] like rotation, scaling etc.

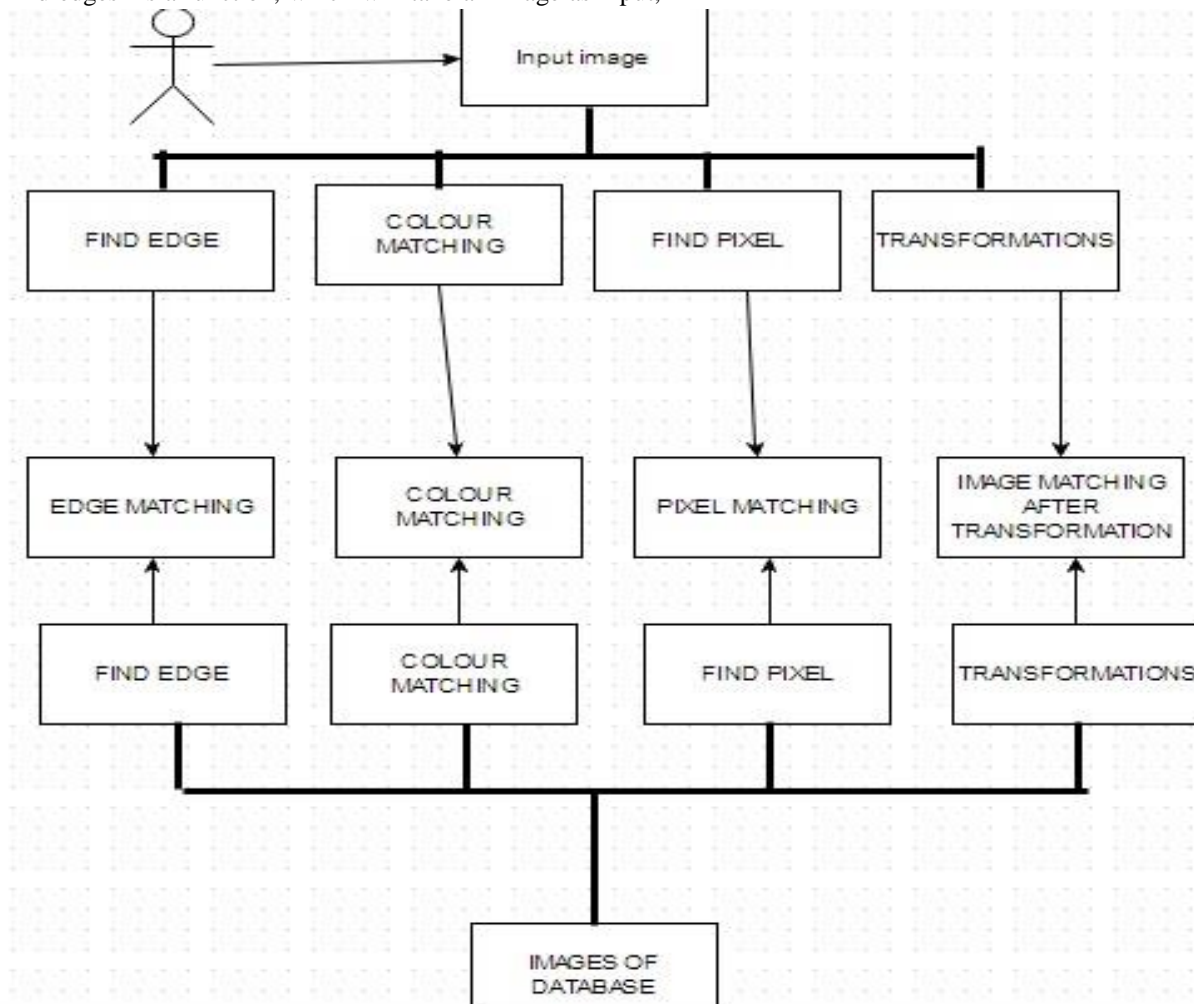


Fig. 1. Flowchart for image search algorithm

A. STEP ALGORITHM FOR EDGE MATCHING:

- 1) Upload image
- 2) Find edges of the image by using "find edges" algorithm.
- 3) Save the image with a name.
- 4) Save the image containing the edge into an array of pixels.
- 5) Match each pixel of the above obtained array in RGB [17] configuration of images in the database and find their difference and percentage difference

B. STEP ALGORITHM FOR CONOUR MATCHING:

- 1) Upload the image
- 2) Find edges of the image by using "find CONTOUR" algorithm.

- 3) Save the image with a name.
- 4) Save the image containing the CONTOUR into an array of pixels.
- 5) Match each pixel in RGB configuration of the images in the database and find their difference and percentage difference and also show the difference graphically.

C. STEP ALGORITHM FOR PIXEL MATCHING:

- 1) Upload the image
- 2) Find the pixel array of the colored version.
- 3) Save the image containing the CONTOUR into an array of pixels.
- 4) Math each pixel in RGB configuration and find the difference and percentage difference and also show the difference graphically.

D. STEP ALGORITHM FOR PIXEL TRANSFORMATION:

1) If the user wants to find a matching between his/her uploaded image and any image in the database where he or she has performed transformation in the uploaded image.

E. OUR APPROACH

In this approach, we follow several steps one after another each of which is explained after the steps.

- User can upload any image of his or her choice.
- Detect the edge of the image.
- Store the RGB (Red, Green, Blue) configuration pixel by pixel for each pixel, in an array. Let us call it pixel array, say pa1.
- Repeat step 2 and 3 for each image stored in the database.
- Now, we compare each pixel array ,i.e pa2, pa3 , pa4 with the pixel array of the uploaded image and find the difference and the percentage unmatched.
- Based on the percentage unmatched, we find out if the images are similar or not.
- We show the difference in a graph.
- Thus we find which images match the most and display them.

Each step is described below in detail:

STEP 1

User can upload any image of his or her choice.



Fig. 2. User can upload any image



Fig. 3. Input image

STEP 2

Next, we detect the edge of the image.

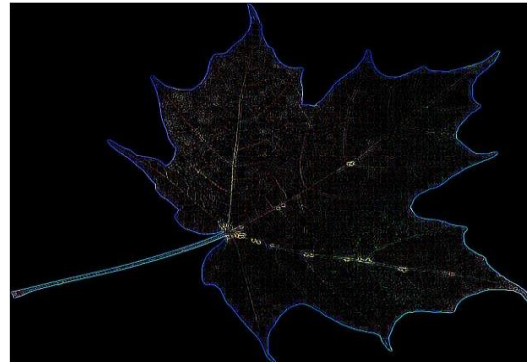


Fig. 4. Image result after Find edges

STEP 3

Next, we find the RGB (Red, Green, Blue) configuration pixel by pixel for each pixel and store it in an array. Let us call it pixel array, say pa1.

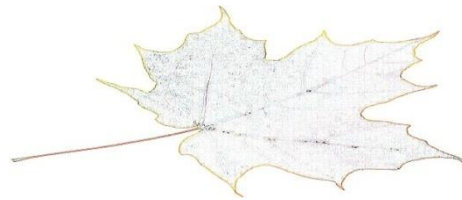


Fig. 5. Result image after contour in found

STEP 4

Repeat step 2 and 3 for each image stored in the database, that is we find out the edge and pixel arrays for each image. Let us name the pixel arrays as pa2, pa3, pa4 and so on.

STEP 5

Now, we compare each pixel array ,i.e pa2, pa3 , pa4 with the pixel array of the uploaded image and find the difference and the percentage unmatched.

STEP 6

Now based on the percentage unmatched, we determine if the images are similar or not. If percentage unmatched is

- less than 20% , then the images are considered to be similar
- greater than 20% but less than 60% , then we rotate the image and find any image matching the transformed, i.e rotated image.
- Greater than 60% , then the images are said to be Unmatched.

STEP 7

Next, we find out the difference between first 20 pixels of two images and plot a graph against pixel number and assume the same pattern to continue for the rest of the pixels.

STEP 8

Thus we find which images match the most, that is find out the images for which the percentage unmatched is the least, and display those images.

V. SCREENSHOTS



Fig. 6. GUI of the application developed



Fig. 7. Output Result images. Percentage differences with input images are also shown for Fig 3 as input.

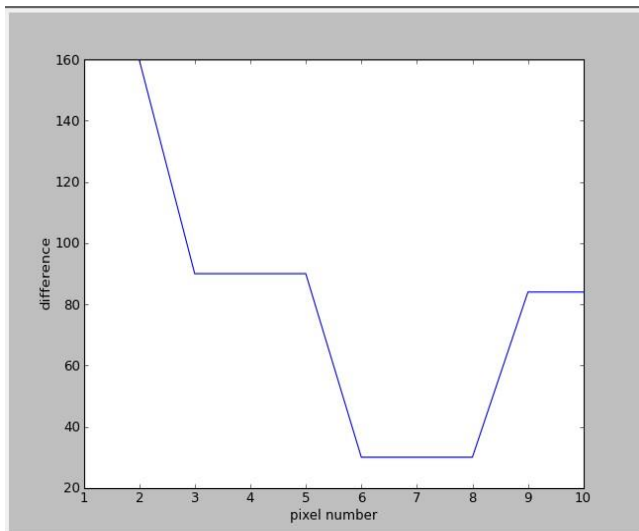


Fig. 8. Graph showing the differences between the two images.

X AXIS: Pixel Number
Y AXIS: Difference.



Fig. 9. Input Image



Fig 10(a)

Fig 10(b)



Fig 10 (c)



Fig 10 (d)

Fig. 10. (a),10(b),10(c),10(d):Output Images of the input image in Fig (9).Input image does not lie in the database. But the rotated versions are present. They are searched and displayed.








Fig. 10. (e)Scaled down version of the input image (Fig 9) occurs in the database and it is displayed



Fig. 10. (f) : Scaled up version of the input image (Fig 9) occurs in the database and it is displayed

TABLE I. PERCENTAGE DIFFERENCE BETWEEN INPUT IMAGE AND DATABASE IMAGES

| | | DATABASE IMAGES | | |
|--------------------------------|---|---|---|--|
| | |  |  |  |
| INPUT I M A G E |  | -44.13 | -28.25 | -54.28 |
| |  | -28.25 | -44.13 | -29.56 |

```

IDLE 2.6.1          ===== No Subprocess =====
>>>
image format: JPEG
image mode: RGB
image size: (272, 185)
image format: PNG
image mode: RGB
image size: (259, 194)
[255 150 200]
3
[255 255 255]
185
ok
unmatched: 25236
102675
25236666
percentage unmatched 24
75710
160
160
90
90
90
30
30
30
84
84

```

Fig. 11. Snapshot of output where percentage differences between input and output images are shown

VI. EXPERIMENTAL EVALUATION

Table 1 shows the percentage differences between the input image and the database images.

VII. DISCUSSION AND CONCLUSIONS

In this study, we have developed an algorithm for shape based image retrieval and image search. We have used an approach where an user uploads an image and first edge detection is done, contour matching is done after contour detection, next pixels are found and stored in an array. Similar steps are performed on database images and percentage differences are found and images are displayed.

Future work will be to improve the algorithm so that the skeleton of images can be found, for example finding leaf skeletons will help in leaf categorizations.

REFERENCES

- [1] www.www2008.org/papers/pdf/p307-jingA.pdf
- [2] <http://infolab.stanford.edu/~wangz/project/imsearch/review/JOUR/>
- [3] en.wikipedia.org/wiki/Syntactic_pattern_recognition
- [4] en.wikipedia.org/wiki/Computer_vision
- [5] www.cs.berkeley.edu/~arbelaez/publications/amfm_pami2011.pdf
- [6] https://vision.in.tum.de/_media/spezial/bib/rosenhahn_iwcia06.pdf
- [7] www.geocomputation.org/1998/99/gc_99.htm
- [8] www.dsp.toronto.edu/~kostas/.../pub/.../2000-SpringerMonograph.pdf
- [9] <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1390657/>
- [10] www.springer.com/?SGWID=0-102-1297-33673609-0
- [11] www.springer.com/?SGWID=0-102-1297-33673609-0
- [12] www.ijarcse.com/docs/papers/Volume_3/11.../V3I11-0501.pdf
- [13] www.math.uci.edu/icamp/summer/research_11/bhonsle/cdfd.pdf
- [14] en.wikipedia.org/wiki/Orbital_eccentricity
- [15] www.mathworks.in/help/matlab/ref/rose.html
- [16] www.ascilite.org.au/ajet/ajet28/guven.pdf
- [17] en.wikipedia.org/wiki/RGB_color_model

Computer Ethics in the Semantic Web Age

Aziz Alotaibi

Department of Computer Science
221 University Ave, University of Bridgeport
Bridgeport, CT, USA

Abstract—Computer ethics can be defined as a set of moral principles that monitor the use of computers. Similar rules were then required for both programmers and users. Issues that were not anticipated in the past have arisen due to the introduction of newer platforms such as Semantic Web. Both programmers and users are now obliged to consider phenomenon such as informed consent. In this paper, I will explore the ethical problems that arise for professionals and users with the advent of new technologies, especially with privacy concerns and global information.

Keyword—Computer ethics; semantic web; privacy concerns; and global information

I. INTRODUCTION

In early 1940s, the computer ethics field was founded by MIT professor Norbert Wiener as academic field[1]. With advancements made in computer science, web technologies provided us with the opportunity to explore a whole new world that we were previously unaware of due to our limited access to resources. However, this knowledge came at a cost[2]. Individuals can no longer be assured that their privacy is maintained on the web. First, people's privacy is being threatened by hackers and computer crime since their information is being controlled by other entities, such as government agencies or other private enterprises.

Nowadays, people have share personal information at a much greater scale than in the past; for instance, their lifestyle, information about their health, their financial status, their political views, their religious affiliation, and gender. Thus, there is greater paranoia in their minds that their information might be accessed by computer hackers. From the hacker's perspective, there should be no restriction to accessing people's information. However, that is not the case and information is being strictly controlled through various means to protect it from being easily shared. Yet, in the new age world, where information our desire to gather information has reached an all-time high, we pay for our quicker, faster access to services by loosening our grip on our privacy and security on the web.

II. SEMANTIC WEB

In 1989, Tim Berners-Lee, at CERN (European Organization for Nuclear Research), made a proposal for a system that would enable people to share the information using the hypertext system. And in 1990, he wrote the first document using HTML[3]. As we know WWW is an abbreviation for World Wide Web which known as (web). And he invented the HTML and HTTP protocol. HTML, stands for hypertext markup language, is used to create and

structure documents that can be shown as web pages. HTTP, short for the hypertext transfer protocol, is an application protocol that use to request and transmit files over the internet. And he invented the HTML and HTTP protocol. HTML, stands for hypertext markup language, is used to create and structure documents that can be shown as web pages[4]. HTTP, short for the hypertext transfer protocol, is an application protocol that use to request and transmit files over the internet. Semantic web is known as web 3.0 and an extension to the World Wide Web that enables to share the data.

As Tim Berners-Lee states semantic web is the movement from the web of documents to the web of data [5]and from the human centric to the machine-processable.

TABLE I. WEB GENERATION

| Document web 1.0 | Social web 2.0 | Semantic web 3.0 | Ubiquitous web 4.0 |
|---|--|--|---|
| Document – centric HTML, CSS, Javascript Search engines URL | Data-centric API, web services, AJAX Social network URI | Data-centric RDF, RDFa, OWL, SPARQL Structure data Linked data | User centric Invisible web Smart market Semantic mobile services |

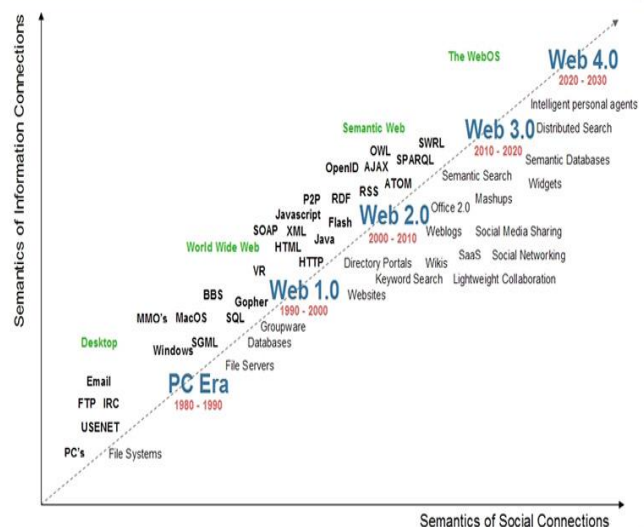


Fig. 1. Source: Radar Networks & Spivack,2007 www.radarnetworks.com

Since these privacy issues already existed prior to the development of newer technologies such as semantic web, its creators and users may question how the introduction of this new platform will affect an already existing problem, i.e., how

will the addition and usage of semantic web further intensify privacy control issues. Information gathering and transaction was possible on the web prior to the discovery of semantic web. However, it is essential to understand that the creation of semantic web has added and made the possibility of data sharing significantly easier than in the past.

III. LINKED DATA

Linked Data is a structure data that has been published and connected by linking different sources over the web. Before semantic web technology has been used, we use the hypertext links to link documents and browse them on the web. Also we use XML to structure the document. Here is an example of XML:

```
<Website>
  <creator name="Aziz">
    <age> 24</age>
  </creator>
  <uri>
    www.unethicalwebsite.com
  </uri>
  <contributor>
    www.yourwebsite.com
  </contributor>
</Website>
```

It is easy to understand XML file, However, XML cannot provide any relationship between the data. Moreover, we cannot query XML file to provide us with an answer.

Semantic web has many languages, and one of them is RDF which is a graph data format. Semantic web uses RDF (Resource Description framework) technology to link the data. RDF can make a relationship between the data over the World Wide Web[4]. The RDF technology relies on URI (uniform Resource Identifier) to link and describe the data items by providing properties and property values as shown in figure 2[6].

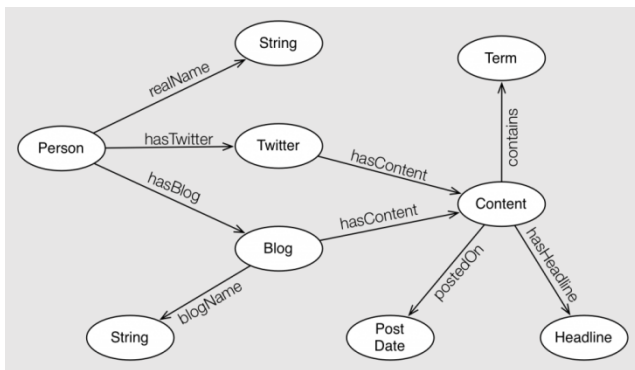


Fig. 2. Linked data

Since these privacy issues already existed prior to the development of newer technologies such as semantic web, its creators and users may question how the introduction of this new platform will affect an already existing problem, i.e., how will the addition and usage of semantic web further intensify privacy control issues. Information gathering and transaction was possible on the web prior to the discovery of semantic web. However, it is essential to understand that the creation of

semantic web has added and made the possibility of data sharing significantly easier than in the past.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:dc="http://purl.org/dc/elements/1.1/" >
  <rdf:Description
    rdf:about="http://www.unethicalwebsite.com">
    <dc:title> unethical websites </dc:title>
  </rdf:Description>
```

```
In this case, anyone can link you to the website as Contributor.
  <rdf:Description rdf:about="http://www.unethicalwebsite.com">
    <dc:Contributor rdf:resource="http://www.yourwebsite.com"/>
  </rdf:Description>
</rdf:RDF>
```

As explained by Shabajee, semantic web operates on the idea of connecting data and metadata in the easiest possible manner. This is made possible by enabling all data to be viewed through a standard format which was a major obstacle in data sharing in the past. Now, with the option of sharing data through semantic web, it is much easier to violate privacy laws unless stricter measures are put in place.

Another function of semantic web is assuring reusability of the data so as to increase its chances of redistribution which leads to its existence even if the original source is removed from the web at a later stage. Therefore, it is highly likely that even if a user deletes specific content from the original location, several other copies would be easily available as a result of prior distribution of the material. This is an example of the effect of new technology on privacy measures and it calls specific checks in place to maintain privacy on the web. The need to find appropriate measures of regulation is critical, as it is possible to immediately integrate the semantic web version of a statement and represent it in results from queries across the whole web. This poses a definite problem since individuals tend to have a variety of information on the web and prefer to separate their professional information from their personal lives. However, the availability to link data through semantic web would connect results from either aspects of an individual's life and can be misused and cause serious damage to their professional reputation or even make them easy victims of stalkers.

Moreover, people get offended on the internet when criticized for their beliefs and their cultural differences. People hold different beliefs and come from societies; they express their point of view on the internet, which sometimes is considered to be offensive by others. People have strong feelings about a particular issue and in the process of expressing their opinion on the web; they end up unintentionally offending those with a different viewpoint on the subject[1]. Thus, if global interaction has to prosper, we are in dire need of educating our users about social rules and expectations so that they can make informed and ethical decisions. This can be achieved by specific long term education process or through publicity or educational campaigns. Educating users about social media ethics would be a step in the right direction so that they are able to communicate with each other in a fair and respectable manner online.

This spread of knowledge should not be limited to just users, web designers and developers should be aware of global information ethics. Moreover, existing design and interface options should be further developed to provide consumers with the level of security that they would feel comfortable using. For instance, giving users a general understanding of how their accounts may be linked will prove beneficial in the long run and may prompt them to read the privacy consent with greater caution before granting their approval. As of now, users have inaccurate ideas of the functionality of most web applications and by adopting models that are easy to understand, we can hope for a more informed population that accesses the web in the future and that in turn will help them to make ethical choices in the electronic world as well.

IV. CONCLUSION

Since the development of computers and the various applications that followed, computer ethics has reached a new found importance due to the emergence of new complications that have accompanied our progress in this field. Traditional web based applications were once considered to be the peak of our grasp over technology. However, human interaction has

achieved a new and rapidly evolving form due to the rise of social media and networking which has redefined our understanding of privacy, made us rethink our control over granting consent and altered the means of global information sharing. The recent breakthroughs in technology such as semantic web are indeed a welcome addition; yet they come with the added responsibility on the part of their developers of ensuring privacy and educating both users and themselves about making ethical choices.

REFERENCES

- [1] Bynum, T.W. and S. Rogerson, Computer ethics and professional responsibility. 2003.
- [2] Hann, I.H., et al. Analyzing Online Information Privacy Concerns: An Information Processing Theory Approach. in System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on. 2007.
- [3] Guha, R. Toward the Intelligent Web Systems. in Computational Intelligence, Communication Systems and Networks, 2009. CICSYN '09. First International Conference on. 2009.
- [4] Bizer, C., The Emerging Web of Linked Data. Intelligent Systems, IEEE, 2009. 24(5): p. 87-92.
- [5] <http://www.w3.org/2001/sw/>, semantic web.
- [6] <http://www.oraclealchemist.com/news/linked-data-rdf-and-sparql-part-1/>.

A Tool Design of Cobit Roadmap Implementation

¹Karim Youssfi, ²Jaouad Boutahar, ³Souhail Elghazi

École Hassania des Travaux Publics
Casablanca, Morocco

Abstract—Over the last two decades, the role of information technology in organizations has changed from primarily a supportive and transactional function to being an essential prerequisite for strategic value generation. The organizations based their operational services through its Information Systems (IS) that need to be managed, controlled and monitored constantly. IT governance (ITG), i.e. the way organizations manage IT resources, has become a key factor for enterprise success due to the increasing enterprise dependency on IT solutions. There are several approaches available to deal with ITG. These methods are diverse, and in some cases, long and complicated to implement. One well-accepted ITG framework is COBIT, designed for a global approach. This paper describes a design of a tool for COBIT roadmap implementation. The model is being developed in the course of ongoing PhD research.

Keywords—IT governance; COBIT; Tool design; Roadmap; Implementation

I. INTRODUCTION

In recent years, due to the increase of IT investment, the IT governance has become a center of interest among practitioners and researchers.

Several issues made its contribution to explain this phenomenon [1]: (1) Business activities became largely dependent in IT systems. (2) Therefore business failure and success are increasingly dependent on IT (3) IT should deliver value to business and be aligned with the organization's goals. (5) Response to fast changes in business environment. (6) Ensure business continuity.

Some methods to support IT governance exist. Weill & Ross have developed an IT governance framework that can be used to assign responsibilities for high level IT decision making, but their work give no more information on how the IT organization must effectively perform their work [2]. The ISO / IEC 20000 and preceding IT Infrastructure Library (ITIL) might aid the creation of processes related to delivery and support [3]. The most recognized, publicly available, IT governance framework is COBIT – Control Objectives for Related Technology– [4], which will be discussed.

These frameworks and standards are useful to guide the decisions of managers on the key processes of IT. However, they remain general framework and must be adapted to the organization. Many organizations struggle with implementing and embedding these governance practices into their organizations. Through case and survey research, it will be vital to verify how organizations are adopting and implementing ITG. This last point is essential: that would guide specification phases of implementation of ITG, reduce costs and deadlines, ensure effective support to implement IT

governance and reduce the risk of failing financial investments. It will be also interesting to analyze this issue in relation to a largely well-accepted framework as COBIT - currently in its fifth edition- covering the IT activities of the enterprise end to end.

Some specific questions are:

- Which COBIT 5 processes and related practices are most adapted to my organization?
- Which COBIT 5 processes and related practices/structures will be easy / difficult to implement?
- How could I implement COBIT 5 processes in my organization?

As a response, this paper proposes to provide a tool design of COBIT roadmap implementation. This paper is organized as follows: Section 2 introduces an overview of IT Governance concepts. Afterward; to encompass the research scope; COBIT 5 framework, its implementation life cycle and available implementation tools will be presented. Then, in section 3, a tool design of COBIT roadmap implementation will be proposed. This paper concludes with discussion and future research directions.

II. LITERATURE REVIEW

A. Information Technology Governance

There are many definitions of Information Technology Governance (ITG)[5], ITG is commonly used to a set of structures and processes to ensure that IT support and adequately maximize the business objectives and strategies of the organization, adding value to the services delivered, weigh the risks and getting a return on investment in IT [5]. The IT Governance is part of a Corporative Governance [6].

In the last decade, the concept of IT governance has attracted the attention among researchers. Those include Brown and Grant [8]; Mähring [9]; Webb, Pollard and Ridley [5]; and Wilkin and Chenhall [11]: (1) Brown and Grant [8] identified three ITG research streams, structural analysis, contingency analysis and the combination of the first two. They contribute a conceptual map of ITG knowledge from literatures. (2) Mähring [9] reviewed ITG literatures that relate to board of directors' role. The study argues that SOX have added compliance pressure and changed board responsibilities. (3) Webb, et al. [5] reviewed a wide range of ITG literatures to integrate [5] presented the diversification and confusion in ITG conceptualization. That review analyzed not only academic but also practical concepts. (4) Wilkin and Chenhall [11] describe concepts of strategic alignment,

performance measurement, risk management, and value delivery as the most significant enablers of IT governance. They note that broader organizational structures, business processes and technology, and resource capabilities influence the enablers and by extension IT governance.

Many researchers also attempt to propose various ITG models and concepts (e.g. Van Grembergen and De Haes [12], Weill and Ross [10], Brown and Grant [6]).

In the practitioner arena, there are a various versions of frameworks and standards dealing with the ITG: ISO/IEC Standard 38500, ITIL V3, and COBIT, for instance, COBIT has been recognized as the most used framework [7].

Past literature reviews indicate different viewpoints and conceptual diversification in ITG field of studies, essentially, when different research communities differently conceptualize ITG. One outstanding finding is that ITG is constantly evolving. Since there are regular introductions of new concepts, legal requirements, standards and practical frameworks. It is vital not to ignore these changes in order to gain better understanding of ITG field.

COBIT 5, the latest version of COBIT [13] is recently introduced, in this context the next section proposes to explore the IT Governance concepts in COBIT 5.

B. IT Governance Concepts in COBIT 5

COBIT is the framework for governance and management of IT developed by ISACA, which evolved into the current version "COBIT 5"- released in 2012, designed to be a single integrated framework [13]. COBIT 5 defines governance as:

"Governance ensures that stakeholder needs, conditions and options are evaluated to determine balanced, agreed-on enterprise objectives to be achieved; setting direction through prioritization and decision making; and monitoring performance and compliance against agreed-on direction and objectives." [13].

This definition is different from the previous versions of COBIT. It recognizes multiple stakeholders of organizational IT as well as balance of resources distribution while maintain overall firm goals. Second, it explicitly states what activities to do. Third, this no mentions about leadership, structures and processes in the definition [14].

COBIT 5 reveals new conceptual ideas compared to previous versions. COBIT 5 proposes COBIT principles, which guide the governance of IT. The five principles include: Meeting Stakeholder Needs; Covering Enterprise End-to-end; Applying a Single, Integrated Framework; Enabling a Holistic Approach; and separating Governance from Management [14] as in Table I. Principle 1 emphasizes on goal cascade and value creation among different stakeholders who may expect different IT value. Principle 2 exhibits that COBIT does not limit to IT department but it covers entire enterprise. COBIT includes guide for integration to corporate governance for value creation by specifying roles, activities and relationships. Principle 3 indicates that COBIT aims to be the umbrella framework.

COBIT provides an integration guideline to use with other frameworks. Principle 4 shows how ITG components relates and provide a set of critical success factors (they are called enablers). Principle 5 shows that COBIT 5 clearly separate governance and management.

TABLE I. COBIT 5 PRINCIPLES [13]

| Principles |
|--|
| Principle 1 - Meeting Stakeholder Needs |
| Principle 2 - Covering the Enterprise End-to-End |
| Principle 3 - Applying a Single Integrated Framework |
| Principle 4 - Enabling a Holistic Approach |
| Principle 5 - Separating Governance from Management |

These principles demonstrate scope, how-to and objectives of COBIT. They highlight on certain concepts, such as, goal cascade and governance enablers.

From operational point of views, COBIT 5 provides 37 processes in two domains. The governance domain contains five processes while management domain contains 32 processes. These processes are provided as a guideline to practitioners. Fig. 1 shows key governance and management areas and Table II shows COBIT processes.

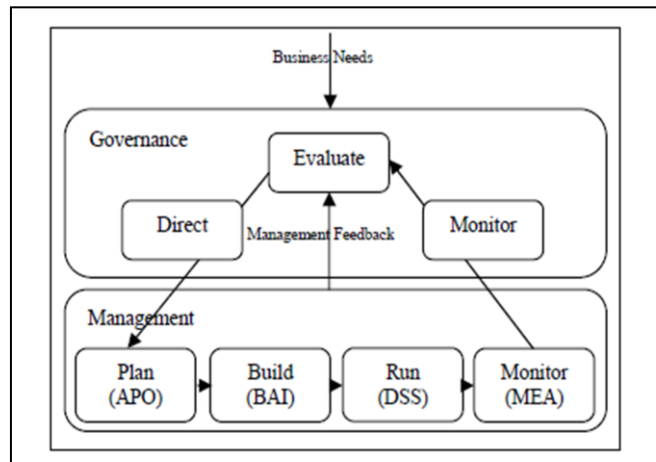


Fig. 1. Cobit 5 Governance and Management Areas [13]

COBIT 5 indicates that governance processes will provide direction to management processes based on business needs.

Then, governance processes will get feedback from management processes to evaluate how well the directions are carried out or whether they are needed to be adjusted.

Governance actions include Evaluate, Direct and Monitor or EDM. COBIT 5 sees board of directors is accountable for governance processes while executives are responsible to perform them. EDM and board accountability concepts are similar to ISO38500 [10].

On the other hand, management processes are categorized by IT life cycle. There are four areas: Align, Plan and Organize (APO); Build, Acquire and Implement (BAI); Deliver, Service and Support (DSS); and Monitor, Evaluate and Assess (MEA).

Each area contains different processes. COBIT 5 sees that APO and MEA areas are directly linked to governance processes. These process areas contain different ITG activities.

COBIT 5 is not a minor update to its previous version. There are conceptual differences, new emphasizes and new arrangements. These distinctions could imply or effect governance practice and knowledge in many ways.

TABLE II. COBIT 5 PROCESS [14]

| Area | Process |
|------|---|
| EDM | EDM1 Set and Maintain the Governance Framework |
| | EDM2 Ensure Value Optimization |
| | EDM3 Ensure Risk Optimization |
| | EDM4 Ensure Resource Optimization |
| | EDM5 Ensure Stakeholder Transparency |
| APO | APO1 Define the Management Framework for IT |
| | APO2 Manage Strategy |
| | APO3 Manage Enterprise Architecture |
| | APO4 Manage Innovation |
| | APO5 Manage Portfolio |
| | APO6 Manage Budget and Cost |
| | APO7 Manage Human Resources |
| | APO8 Manage Relationships |
| | APO9 Manage Service Agreements |
| | APO10 Manage Suppliers |
| | APO11 Manage Quality |
| | APO12 Manage Risk |
| | APO13 Manage Security |
| BAI | BAI1 Manage Programs and Projects |
| | BAI2 Define Requirements |
| | BAI3 Identify and Build Solutions |
| | BAI4 Manage Availability and Capacity |
| | BAI5 Manage Organizational Change Enablement |
| | BAI6 Manage Changes |
| | BAI7 Manage Change Acceptance and Transitioning |
| | BAI8 Manage Knowledge |
| | BAI9 Manage Assets |
| | BAI10 Manage Configuration |
| DSS | DSS1 Manage Operations |
| | DSS2 Manage Service Requests and Incidents |
| | DSS3 Manage Problems |
| | DSS6 Manage Continuity |
| | DSS5 Manage Security Services |
| | DSS6 Manage Business Process Controls |
| MEA | MEA1 MEA Performance and Conformance |
| | MEA2 MEA the System of Internal Control |
| | MEA3 MEA Compliance with External Requirements |

C. COBIT 5 Implementation life cycle

COBIT 5 has a professional guide for implementation. The guide provides details of seven phases of the implementation life cycle, applying a continual improvement life cycle approach provides a method for enterprises to address the complexity and challenges typically encountered during ITG implementation [14]. There are three interrelated dimensions to the life cycle, as illustrated in figure 2: the core ITG continual improvement life cycle, the enablement of change (addressing the behavioral and cultural aspects of the implementation or improvement), and the management of the Program. The three aforementioned dimensions exist within each and every one of these phases

The seven phases of the implementation life cycle are illustrated in figure 2.

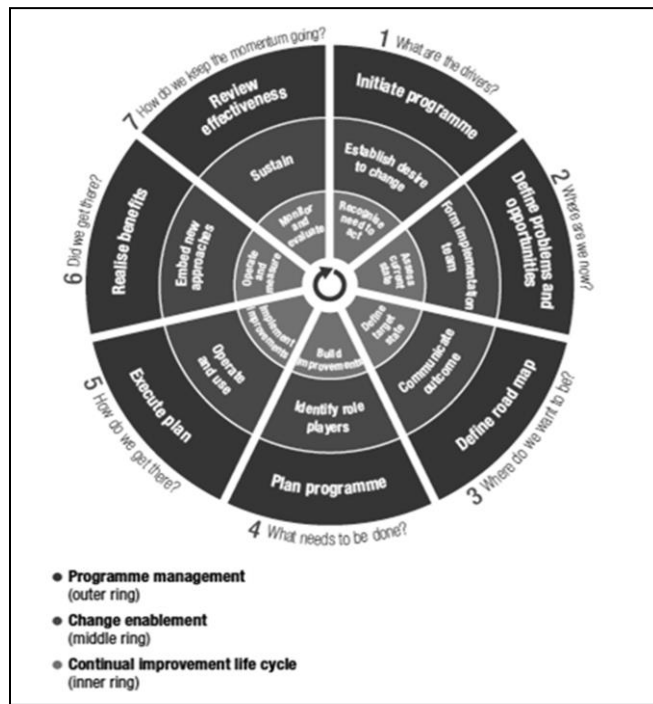


Fig. 2. Seven Phases of the Implementation Life Cycle[14]

Phase 1—What Are the Drivers?

Phase 1 identifies current change drivers and creates at executive management levels a desire to change.

Key Questions, which need to be answered in this phase, include: What is the business motivation and justification? What are the Stakeholder needs and expectations that need to be satisfied? Why are we doing this?

There must be consensus on the need for implementing COBIT 5, to change and improve, supported by the will and commitment of executive management.

Dimensions:

- Program Management – Initiate the Program
- Change Enablement – Establish the desire to change
- Continual Improvement Lifecycle – Recognize the need to act.

Phase 2—Where Are We Now?

Phase 2 aligns IT-related objectives with enterprise strategies and risk, and priorities the most important enterprise goals, IT-related goals and processes. COBIT 5 provides a generic mapping of enterprise goals to IT-related goals to IT processes to help with the selection. Given the selected enterprise and IT-related goals, critical processes are identified that need to be of sufficient capability to ensure successful outcomes. Management needs to know its current capability and where deficiencies may exist. This is achieved by a process capability assessment of the as-is status of the selected processes.

Dimensions:

- Program Management – Define Problems and Opportunities
- Change Enablement – Form the implementation team
- Continual Improvement Lifecycle – Assess current state

Phase 3—Where Do We Want To Be?

Phase 3 sets a target for improvement followed by a gap analysis to identify potential solutions. Some solutions will be quick wins and others more challenging, long-term tasks. Priority should be given to projects that are easier to achieve and likely to give the greatest benefit. Longer-term tasks should be broken down into manageable pieces.

A defined target is set for the future improvement, a gap analysis is completed to indicate the delta between as-Is and To-Be, and potential improvements are identified.

Dimensions:

- Program Management – Define the Roadmap
- Change Enablement – Communicate outcome
- Continual Improvement Lifecycle – Define target state

Phase 4—What Needs To Be Done?

Phase 4 plans feasible and practical solutions by defining projects supported by justifiable business cases and developing a change plan for implementation. A well-developed business case will help ensure that the project’s benefits are identified and continually monitored.

Comprehensive business cases and change plans are developed, and projects planned, for delivering the work and effecting the implementation into the Enterprise.

Dimensions:

- Program Management – Plan Program
- Change Enablement – Identify role players
- Continual Improvement Lifecycle – Build improvements

Phase 5—How Do We Get There?

Phase 5 provides for the implementation of the proposed solutions into day-to-day practices and the establishment of measures and monitoring systems to ensure that business alignment is achieved and performance can be measured.

Success requires engagement, awareness and communication, understanding and commitment of top management, and ownership by the affected business and IT process owners.

Dimensions:

- Program Management – Execute plan
- Change Enablement – Operate and use
- Continual Improvement Lifecycle – Implement improvements

Phase 6—Did We Get There?

Phase 6 focuses on sustainable transition of the improved governance and management practices into normal business operations and monitoring achievement of the improvements using the performance metrics and expected benefits.

Dimensions:

- Program Management – Realize benefits
- Change Enablement – Embed new approaches
- Continual Improvement Lifecycle – Operate and measure

Phase 7—How Do We Keep the Momentum Going?

Phase 7 reviews the overall success of the initiative, identifies further governance or management requirements and reinforces the need for continual improvement. It also priorities further opportunities to improve GEIT.

Dimensions:

- Program Management – Review effectiveness
- Change Enablement – Sustain
- Continual Improvement Lifecycle – Monitor and evaluate

The time spent per phase will differ greatly depending on (amongst other factors) the specific enterprise environment, its maturity, and the scope of the implementation or improvement initiative. However, the overall time spent on each iteration of the life cycle ideally should not exceed six months, with improvements applied progressively; otherwise, there is a risk of losing momentum, focus and buy-in from stakeholders.

Over time, the life cycle will be followed iteratively while building a sustainable approach. This becomes a normal business practice when the phases in the life cycle are everyday activities and continual improvement occurs naturally.

Figure 3 illustrate an example of generic roles for key stakeholders and responsibilities of implementation role players when creating the appropriate environment to sustain governance and ensure successful outcomes. Similar tables are provided for each phase of the implementation life cycle.

| Key Activities | Responsibilities of Implementation Role Players | | | | | | | | |
|--|---|------------------------|-----|--------------------|-------------|-------------------|----------|---------------------|--------------------|
| | Board | IT Executive Committee | CIO | Business Executive | IT Managers | IT Process Owners | IT Asset | Risk and Compliance | Programme Steering |
| Set direction for the programme. | A | R | R | C | C | I | C | C | C |
| Provide programme management resources. | C | A | R | R | C | C | R | R | I |
| Establish and maintain direction and oversight structures and processes. | C | A | C | I | I | I | I | I | R |
| Establish and maintain programme. | I | A | R | C | C | I | I | I | R |
| Align approaches with enterprise approaches. | I | A | R | C | C | I | C | C | R |

A RACI chart identifies who is Responsible, Accountable, Consulted and/or Informed.

Fig. 3. Creating the Appropriate Environment RACI Chart[14]

D. Available tools

In addition, to the implementation guide described in the previous section, there are a number of tools included within the guidance:

1) *Assessment Scoping Tool*—An Excel file that brings together various existing mappings related to COBIT 5 in a hierarchical tree format, including:

- Mapping of COBIT 5 processes to IT goals to business goals to IT balanced scorecard
- Mapping COBIT 5 processes to IT goals

2) *Self-assessment Templates*—An Excel file with separate evaluation sheets for all 37 COBIT 5 processes.

Except for the documentations provided by ISACA to their members, there is a lack of important documentation from other sources regarding the latest version of the framework. For this reason, this paper is based on ISACA documentation.

Our analysis on COBIT 5 implementation guide also reveals that the implementation guidance builds extensively on all the COBIT components such as [14]–[15]–[16], so the team in charge of the IT Governance Implementation should be already familiar with all other COBIT 5 guidance. This multitude and complexity of the guides can be an obstacle for the implementation of COBIT; in this context the next section proposes a tool design of COBIT roadmap implementation.

III. A TOOL DESIGN OF COBIT ROADMAP IMPLEMENTATION

COBIT is a largely well-accepted ITG framework; COBIT5 the last version of COBIT offers a wide range of guides (COBIT5: Process facilitating, for implementation, for information security...)

For COBIT 5 implementation, ISACA suggests a lifecycle approach based on 7 phases with high-level roles. However, the multitude and complexity of the guides can be an obstacle for the implementation of COBIT; as a solution to these issues, we propose a tool design of COBIT roadmap implementation such tool would ensure effective support to enterprises wishing to implement COBIT.

In COBIT 5 implementation Guide [14], ISACA propose a lifecycle of 7 phases, our tool will support the first 4 phases in the COBIT implementation life cycle that deal with the establishment of a roadmap of COBIT implementation:

- Initiate Program
- Define problems and opportunities
- Define Roadmap
- Plan Program

The RACI matrices provided by COBIT states that each implementation related activity might be associated with a role, so that the role is responsible, accountable, consulted or informed with respect to the activity.

Implementation Guide introduces 9 different stakeholders. Our proposal features a more simplified representation of only 5 different stakeholders by considering that consulted or informed stakeholders are inactive.

TABLE III. ROLES FOR COBIT ROADMAP IMPLEMENTATION TOOL

| Role | Description |
|--------------------------------|---|
| Program Steering | Direct, design, control, drive and execute the end-to-end Program from the identification of objectives and requirements, to the eventual evaluation of business case objectives and the identification of triggers and objectives for implementation or improvement cycles. |
| Assessment Responsible | Participate as required throughout the Program and provide assessment inputs on relevant issues. Plan, perform and verify assessment results independently. Provide advice on current issues being experienced and input on control practices and approaches. Review the feasibility of business cases and implementation plans. Provide guidance as required during implementation. |
| CEO | Provide leadership to the Program and applicable IT resources to the core implementation team. Work with business management and executives to set the appropriate objectives, direction and approach for the Program. |
| Business Executive | Provide applicable business resources to the core implementation team. Work with IT to ensure that the outcomes of the improvement Program are aligned to and appropriate for the business environment of the enterprise, and that value is delivered and risk is managed. Visibly support the improvement Program and work with IT to address any issues that are experienced. Ensure that the business is adequately involved during implementation and in the transition to use. |
| Board and executive management | Set the overall direction, context and objectives for the improvement Program and ensure alignment with the enterprise business strategy, governance and risk management. Provide visible support and commitment for the initiative, including the roles of sponsoring and promoting the initiative. Approve the outcomes of the Program, and ensure that envisioned benefits are attained and corrective measures are taken as appropriate. Ensure that the required resources (financial, human and other) are available to the initiative. |

Given that an IT organization desires to move from a current state, the as-is model, through evaluating a number of possible change scenarios, to the desired to-be scenario, seven steps needs to be taken, Figure 9 provides a BPMN modeling of steps cited below:

1) *Define scope*: The COBIT framework is a general framework, suitable for many different types of enterprises, as discussed. In order to align effort with the real needs of the enterprise, the roadmap begins with establishing clear goals among the generic COBIT enterprise goals distributed according Balance Score Card four dimensions (Financial, Customer, Internal, Learning/Growth).

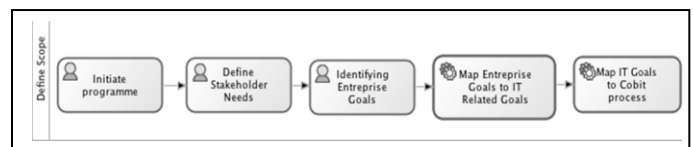


Fig. 4. Overview of define Scope steps

COBIT provides goals cascade to translate stakeholder needs into specific, actionable and customized enterprise goals and into IT related goals. COBIT provides also a mapping between IT-related goals and the relevant COBIT processes. When this logical sequence is followed, the system can deduce the IT processes to implement or improve.

2) *Create As-is Model of Current IT Organization:* The second step concerns the development of a model of the current IT organization.

In order to assess the maturity, an as-is model of Current IT Organization is created based on structure of COBIT 5 Process Reference Model (PRM) defined in Process Assessment Model: Using COBIT 5. The reference model is a predefined, optimal IT governance model that represents the ideal organization, COBIT 5 PRM subdivides the IT-related processes, practices and activities of the enterprise into two main areas, governance and management. Governance ensures that stakeholders needs, conditions and options are evaluated to determine balanced, agreed-upon enterprise objectives to be achieved, setting direction through prioritization and decision making, and monitoring performance and compliance against enterprise objectives. Management ensures that the plan, build, run and monitor (PBRM) IT management activities are executed in alignment with the direction set by the governance body to achieve the enterprise objectives.

COBIT describes a PRM in term of:

- Purpose
- Outcomes
- Base Practices: the activity needed to accomplish the process outcome.
- Input and Output Work products.

By using such Model, it is possible to create a model of current IT organization's governance structure.

3) *Assess current maturity level:* The third step is to assess the capability level of a process ("as-it maturity").

The Capability Model is based on ISO/IEC 15504 (SPICE):

- Level 0: Incomplete. The process is not implemented or fails to achieve its purpose;
- Level 1: Performed (Informed). The process is implemented and achieves its purpose;
- Level 2: Managed (Planned and monitored). The process is managed and results are specified, controlled and maintained;
- Level 3: Established (Well defined). A standard process is defined and used throughout the organization;
- Level 4: Predictable (Quantitatively managed). The process is executed consistently within defined limits
- Level 5: Optimizing (Continuous improvement). The process is continuously improved to meet relevant current and projected business goals.

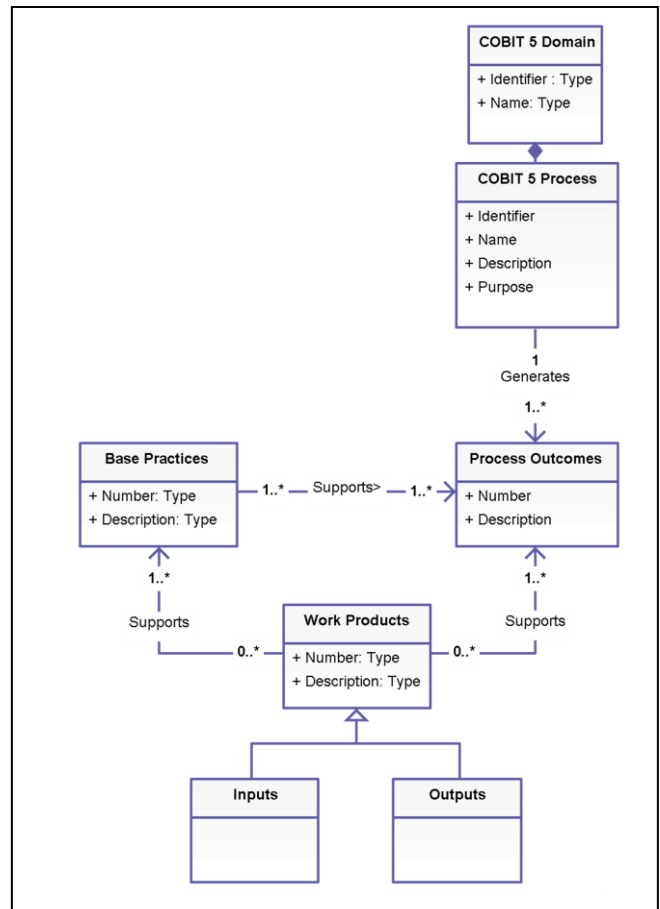


Fig. 5. Class Diagram of COBIT 5 Process Reference Model

The capability of processes is measured using process attributes. The international standard defines nine process attributes [15]:

- 1.1 Process Performance
 - 2.1 Performance Management
 - 2.2 Work Product Management
- 3.1 Process Definition
 - 3.2 Process Deployment
- 4.1 Process Measurement
 - 4.2 Process Control
- 5.1 Process Innovation
 - 5.2 Process Optimization.

Each process attribute is assessed on a four-point (N-P-L-F) rating scale:

- Not achieved (0 - 15%)
- Partially achieved (>15% - 50%)
- Largely achieved (>50% - 85%)
- Fully achieved (>85% - 100%)

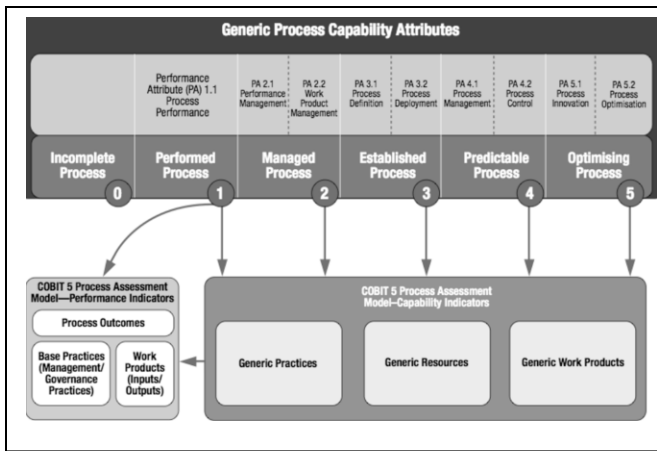


Fig. 6. COBIT 5 Process Capability Model[15]

In COBIT 5 to achieve a given level of capability, the previous level has to be completely achieved.

The maturity level will be the result of comparison between as-is Model of Current IT Organization and the COBIT PRM. Figure 7 shows an overview of assessment method.

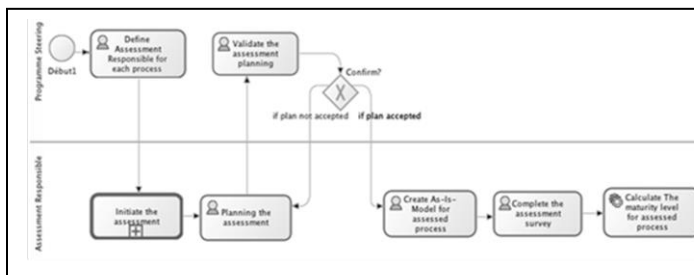


Fig. 7. Overview of Assessment Method

4) *Identify potential Change Scenarios:* In order to identify the potential improvements, IT managers and business managers are interviewed to establish the To-Be maturity level based on enterprise requirement for performance and conformance, the reasons for not achieving this level can be calculated from the approach explained above, and potential improvements can be defined:

The system identify performs a comparison (by attribute) between current capability model and target capability level.

a) *If a required process outcome is not consistently achieved, the process does not meet its objective and needs to be improved.*

b) *The assessment of the process practices will reveal which practices are lacking or failing, enabling implementation and/or improvement of those practices to take place and allowing all process outcomes to be achieved.*

Once gaps identified, Program steering can define potential improvements:

- Collate Gaps into potential improvements.
- Prioritize and argue every potential improvement.

5) *Prioritize and select change scenarios:* Decision-making can be described as a process of improvement selection. For each improvement, the decision-maker should consider the potential benefit, ease of implementation (cost, effort, sustainability), and risk.

Unapproved projects and initiatives should also be recorded for potential future consideration.

6) *Establish the roadmap:*

The approved improvements should be integrated into an overall improvement strategy with a detailed plan to roll out the solution.

This step consists of:

- Defining and gather approved improvements into projects needed to implement the To-Be scenario.
- Developing a Program plan with allocated resources and project plans, and defines the projects deliverables.
- Identify metrics for measuring the progress.

Figure 8 shows the use case diagram of COBIT roadmap implementation tool.

IV. CONCLUSION

This paper has presented a tool design of COBIT roadmap implementation; our design was based mainly on the COBIT 5 lifecycle of implementation. The purpose of such tool is to industrialize the setup of COBIT; reduce costs and deadlines; ensure guidance and effective support through the IT governance implementation life cycle phases; and reduce the risk of failing financial investments.

Further, because the lifecycle presented in COBIT 5 implementation guide provides only generic guidance, the IT governance implementation roadmap is not prescriptive and should be tailored to the needs of the organization applying it. The tool will provide an efficient method for implementing IT governance using COBIT 5 and adapt the roadmap to the effective need of the organization.

V. FUTURE RESEARCH DIRECTION

Further research is ongoing to provide a set of key indicators in order to give a widespread support decision-making in the selection and prioritization change scenarios. The implementation guide describes briefly some indicators such potential benefit, ease of implementation (cost, effort, sustainability), and risk; other economic and financial indicators like value creation, and ROI will be considered as evaluation variables.

COBIT 5 management practices, and Other Specific frameworks: such as PMBOK, can also provide guidance through for this step.

In the next step, implementation phase will be started; as envisaged in the design science research paradigm [16], an evaluation of the tool will be also performed:

- In a first step multiple explorative focus groups will be used to evaluate the perceived utility and actual usability of the developed tool.

- Secondly, laboratory experiments will be carried out to quantitatively measure the effectiveness to validate if the usage of the proposed tool will reduce the perceived complexity costs and deadlines of COBIT 5 implementation phases.

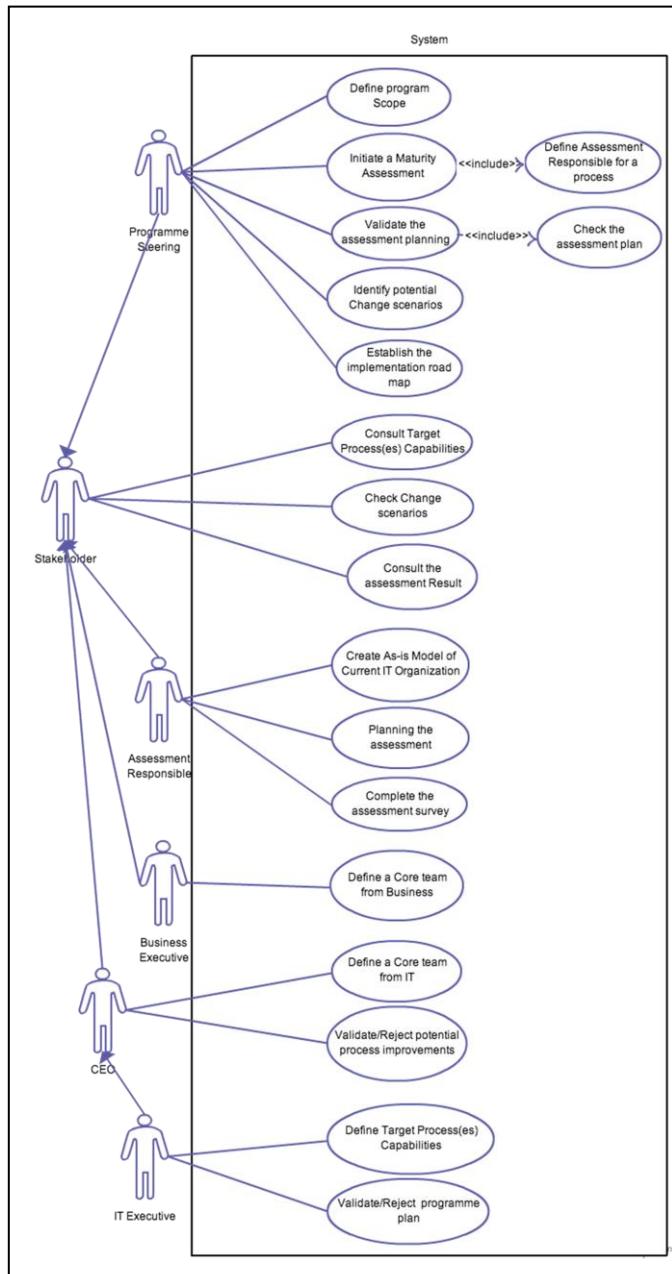


Fig. 8. Use case diagram of COBIT roadmap implementation tool

REFERENCES

- [1] JoséTomaz, "Using Agent-based Simulation for IT Governance Evaluation", Proceedings of the 45th HICSS, Doctoral Program on Complexity Sciences, June 15, 2011, pp. 1.
- [2] Weill, P. and J.W. Ross, "IT governance – How Top Performers Manage IT Decision Rights for Superior Results", Harvard Business School Press, 2004.
- [3] International Organization for Standardization, ISO/IEC 20000-1 & ISO/IEC 20000-2, 2005.
- [4] Information Systems Audit and Control Association, Control Objectives for Information and Related Technology, 4th Edition, 2005.
- [5] P. Webb, C. Pollard, and G. Ridley, "Attempting to Define IT Governance: Wisdom or Folly?," Hawaii International Conference on System Sciences, 2006.
- [6] G. Grant, A. Brown, A. Uruthirapathy, and S. McKnight, "An Extended Model of IT Governance: A Conceptual Proposal", AMCIS 2007, 2007, p. 215.
- [7] Price waterhouse Coopers, "IT Governance Global Status Report," 2008.
- [8] A. E. Brown and G. G. Grant, "Framing the Frameworks: A Review of IT Governance Research," Communications of the Association for Information Systems, vol. 15, pp. 696-712, 2005.
- [9] M. Mähring, "The Role of the Board of Directors in IT Governance: A Review and Agenda for Research," in AMCIS 2006, 2006, p. 377.
- [10] P. Weill, and J. Ross, "IT governance: how top performers manage IT decision rights for superior results," Harvard Business School Press, 2004.
- [11] C. L. Wilkin and R. H. Chenhall, "A Review of IT Governance: A Taxonomy to Inform Accounting Information Systems", Journal of Information Systems, vol. 24, pp. 107-146, 2010.
- [12] W. Van Grembergen and S. De Haes, Enterprise governance of information technology: achieving strategic alignment and value, Springer-Verlag New York Inc, 2009.
- [13] ISACA, COBIT 5: A Business Framework for the Governance and Management of Enterprise IT, 2012.
- [14] ISACA, COBIT 5: Implementation, 2012.
- [15] ISACA, Assessor Guide: Using COBIT@5, 2013.
- [16] O.Müller, S.Debortoli, S.Seidel, "Implementation of a Design Theory for Systems that Support Convergent and Divergent Thinking," pp. 438-445, 2013.

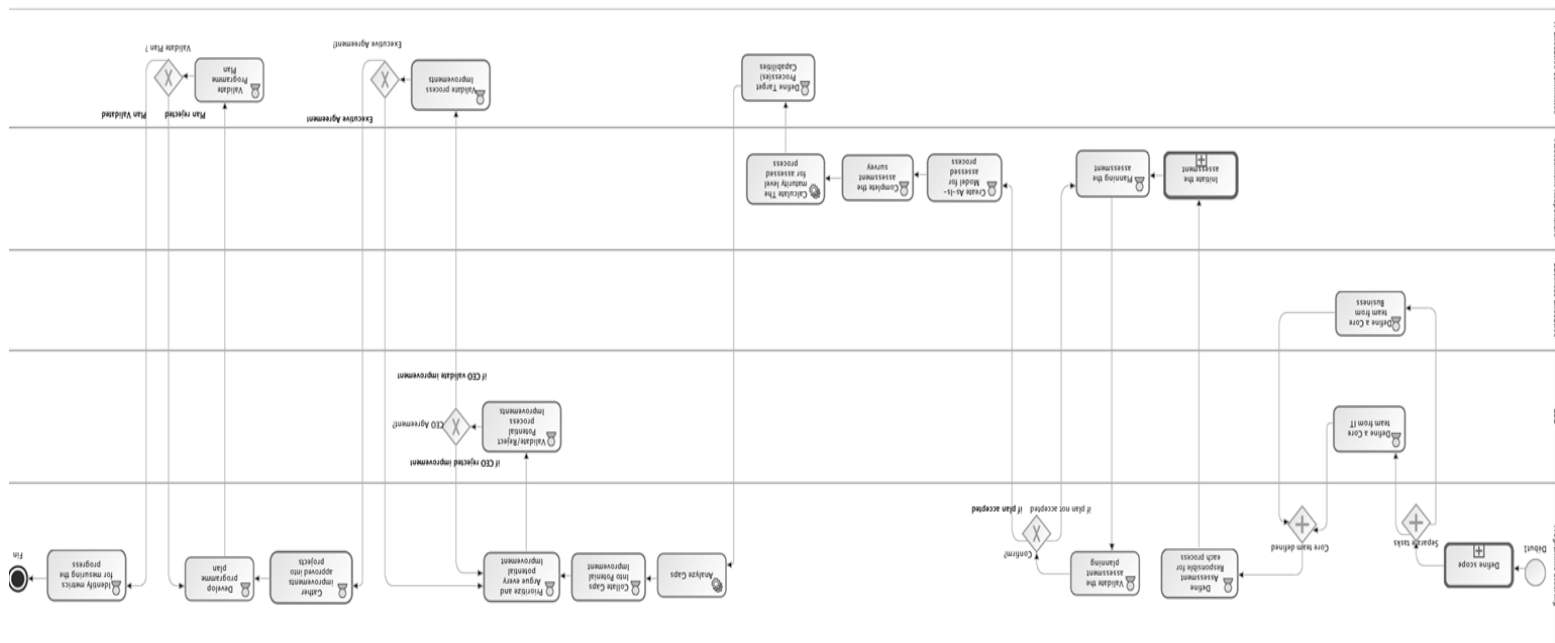


Fig. 9. BPMN Modeling of COBIT roadmap implementation

Ontology Mapping of Business Process Modeling Based on Formal Temporal Logic

Irfan Chishti

Department of Computing and Information Systems
University of Greenwich
London, UK

Jixin Ma¹, Brian Knight²

Department of Computing and Information Systems
University of Greenwich,
London, UK

Abstract—A business process is the combination of a set of activities with logical order and dependence, whose objective is to produce a desired goal. Business process modeling (BPM) using knowledge of the available process modeling techniques enables a common understanding and analysis of a business process. Industry and academics use informal and formal techniques respectively to represent business processes (BP), having the main objective to support an organization. Despite both are aiming at BPM, the techniques used are quite different in their semantics. While carrying out literature research, it has been found that there is no general representation of business process modeling is available that is expressive than the commercial modeling tools and techniques. Therefore, it is primarily conceived to provide an ontology mapping of modeling terms of Business Process Modeling Notation (BPMN), Unified Modeling Language (UML) Activity Diagrams (AD) and Event Driven Process Chains (EPC) to temporal logic. Being a formal system, first order logic assists in thorough understanding of process modeling and its application. However, our contribution is to devise a versatile conceptual categorization of modeling terms/constructs and also formalizing them, based on well accepted business notions, such as action, event, process, connector and flow. It is demonstrated that the new categorization of modeling terms mapped to formal temporal logic, provides the expressive power to subsume business process modeling techniques i.e. BPMN, UML AD and EPC.

Keywords—Business Process Modeling techniques; Ontology; Temporal Logic; Semantics; Mapping

I. INTRODUCTION

Business Process (BP) is defined [16], [24], referring to a structure set of actions designed to show how work is done, rather than what is done. The actions referred to are usually work elements, producing some component or subcomponent of a complete artefact. Actions are structured according to essential logical time ordering of component production. However, there is still a difference between meanings and use of the terms utilized in different modeling techniques. In the business and management field, processes are described mainly for human to human communication, for decision making in production processes, administrative processes, to understand their impact on the organization. In the technical field, processes are considered as a form of high level programming languages, conceived to achieve a better use of web services (and, more generally, e-services), i.e., they represent an executable form of the application logics, as part of a complex software artefact.

Business process models are created “to understand the key mechanisms of an existing business; to orient the creation of suitable information systems that support the business; to implement improvements in the current business; to show the structure of an innovated business; to experiment new business concepts; and to identify business elements not considered part of the core, which could be delegated to an outside supplier” [25]. Hence the job to describe business processes and modeling them is becoming increasingly complex. Both industry experts and academics in the field of business process re-engineering and business process management have concluded that successful systems start with an understanding of the business processes of an organization.

In logic, initial attempts to describe fundamental structures of our world were made by Aristotle. To reason and represent process, temporal systems/frameworks used different objects to represent temporal ontology i.e. interval and moment. In 19th century, Charles Sander Peirce invented classical first order logic, which provides a powerful instrument for representing any factual information. However, it is necessary to ask why it is useful to express processes using first order logic. The reason is that the current literature does not provide a logical foundation of business process modeling and logical knowledge representations are loosely coupled with artificial intelligence.

Formal representations allow for better analysis of the designs to identify the process improvements that can lead to increased profitability and improved productivity. The primary aim of this paper is therefore empirical study of main business process modeling techniques and tools explicitly aimed at modeling business processes with the intent of providing a mapping of modeling terms to temporal logic. To achieve this, a versatile conceptual categorization of commercial modeling terms is proposed which gathers a wide variety of commercial modeling terms into a three distinguishable categories and subsequently formalized them. This will ease the process of ontology mapping and also combine key advantages of commercial modeling techniques by providing rigorous logical basis which is general and expressive enough. Formal representation of the processes also forms the basis for future automation of tasks which make up a business process.

Nevertheless, other techniques might exist that are used or that might be used for modeling business processes, which are not considered in this paper. Such techniques applicable to processes in general are applicable to business processes in particular. However, whether all business process modeling

tools and techniques are applicable or not to process modeling is beyond the scope of this paper. The focus of this paper is BPM tools and techniques such as BPMN, UML AD and EPC and temporal model for business processes [18]. The findings of this paper provide an ontology that is general and expressive enough to subsume commercially accepted modeling terminology and characterized by the following properties:

- Main terms/constructs corresponding to concepts and modeling notions drawn from the BPM techniques.
- Conceptually categorize them based on their dynamic and static properties.
- Formally define the conceptual categories.
- Represent them graphically using logical structure to show its generality, simplification and expressiveness.

The rest of the paper is organized as follows. Section II presents the related work; section III introduces the main BPM techniques i.e. BPMN, UML AD and EPC modeling terms/constructs, and author's contribution; a conceptual categorization of modeling terms and its formal definitions. Section IV discusses about temporal logic/systems and main terms/constructs from temporal model for business processes [18], while section V author's contribution towards an ontology mapping of modeling terms discussed in section III and IV, followed by graphical representation of them using an example and section VI will provide the conclusion and possible future work.

II. RELATED WORK

Logical modeling of business processes is to facilitate the understanding and development of business process model that supports the organization, and to permit the analysis and re-engineering or improvement of them. Even though it was 1960 when Levitt first mentioned the importance of business processes it was not until the 1990s that processes have acquired a real importance in enterprise design [25]. Davenport [24], Hammer and Champy [15], [16], and Harrington [10] have promoted the new perspective. However, increasing popularity of business process orientation has attracted designers and developers in the industry to develop new methodologies, and modeling tools and techniques to support it. The task of describing and modeling business process has become more complex due to increased availability of different modeling techniques with the lack of a guide that explains and describes the logic and concepts involved. In this section we briefly present the major BPM techniques and languages:

A. Business Process Modeling Notation (BPMN)

BPMN is the most recent standard notation proposed by Object Management Group (OMG) to design business processes. The primary goal of *BPMN* is to provide a notation that is readily understandable by all business users, from the business analysts that create the initial drafts of the processes, to the technical developers responsible for implementing the technology that will perform those processes, and finally, to the business people who will manage and monitor those processes.

Thus, *BPMN* creates a standardized bridge for the gap between the business process design and process implementation¹.

B. Unified Modeling Language (UML)

UML is an Object Management Group (OMG) standard which provides the specification for a graphical, general purpose, object-oriented modeling language. *UML* Activity diagrams(*AD*) are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. *UML AD*, are intended to model both computational and organizational processes (i.e. workflows) Activity diagrams show the overall flow of control².

C. Event Driven Process Chains (EPC)

EPC is a modeling technique for business processes modeling developed in the 1990's. Event-driven process chains are an important notation to model the domain aspects of business processes. The main focus of this rather informal notation is on representing domain concepts and processes rather than their formal aspects or their technical realization. Event-driven process chains are part of a holistic modeling approach, called the ARIS framework; ARIS stands for Architecture of Integrated Information Systems, and it was developed by August-Wilhelm Scheer [2].

D. ICAM Definition method (IDEF)

IDEF's roots began when the US Air Force, in response to the identification of the need to improve manufacturing operations, established the Integrated Computer-Aided Manufacturing (ICAM) program in the mid-1970s. The requirement to model activities, data, and dynamic (behavioral) elements of the manufacturing operations resulted in the initial selection of the Structured Analysis and Design Technique (SADT). The Integrated Definition for Function Modeling (IDEF) is a family of methods that supports a paradigm capable of addressing the modeling needs of an enterprise and its business areas. Among these techniques we mention:

- IDEF0: the function modeling method, and
- IDEF3: the process description captures method.

IDEF0 is a method designed to model the decisions, actions, and activities of an organization or a system. It allows activities and important relations between them to be represented in a nontemporal fashion. It does not support the complete specification of a process³. *IDEF3* provides a mechanism for collecting and documenting processes, by capturing precedence and causality relations between situations and events. There are two IDEF3 description modes⁴:

- *process flow*: capturing knowledge of "how things work" in an organization, and

¹ OMG. Business Process Model and Notation (BPMN), 2011
<http://www.omg.org/spec/BPMN/2.0,formal/2011-01-03>.

² Unified Modeling Language: Superstructure version 2.1.1.
(<http://www.omg.org/docs/formal/07-02-03.pdf>).

³ IDEF Function Modeling Method. [<http://www.idef.com/IDEF0.html>]

⁴ IDEF Process Description Capture Method. [<http://www.idef.com/IDEF3.html>]

- *object-state transition network*: summarizing allowable transitions an object may undergo throughout a particular process.

E. Business Process Execution Language for Web Services (BPEL4WS, or BPEL:

BPEL/BPEL4WS is a de-facto standard for implementing processes based on web services. According to BPEL, processes can be described as:

- *Executable processes*: modeling the behavior of a participant in a business interaction, or as
- *Abstract processes*: specifying the mutually visible message exchange among the parties involved in the protocol, without revealing their internal behavior.

To obtain an executable BPEL process, modelers need to specify primitive and structured activities, execution ordering, messages exchanged, and fault and exception handling. Furthermore, a recent proposal, BPEL4People⁵, extends BPEL4WS specification to describe scenarios where users are involved in business processes. BPEL is a powerful and a widely adopted standard. Among its major drawbacks are; its inherent complexity, the verbosity of the XML encoding and the lack of a specific graphical representation. Such characteristics make it scarcely accepted by business people.

F. XML Process Definition Language (XPDL)

XPDL is a Workflow management coalition (WfMC) standard for interchanging process models among process definition tools and workflow management systems. It provides the modeling constructs of BPMN and allows a BPMN process to be specified as an XML document. XPDL process models can be run on compliant execution engines, even if has been originally conceived as a process design and interchange format specifically for BPMN. It represents the linear form of the process definition based on BPMN graphics⁶.

G. Petri Net or place/transition net

Petri Net is one of several mathematical representations of discrete distributed systems [11]. As a modeling language, it graphically depicts the structure of a distributed system as a directed bipartite graph with annotations. A Petri net consists of *places*, *transitions*, and *directed arcs*, where

- Arcs run between places and transitions,
- Places may contain any number of tokens, and
- Transitions act on input tokens by a process known as *firing*.

Execution of Petri nets is nondeterministic. This means two things: multiple transitions can be enabled at the same time,

any one of which can fire, none are *required* to fire — they fire at will, between time 0 and infinity, or not at all. Since firing is nondeterministic, Petri nets are well suited for modeling the concurrent behavior of distributed systems [12].

H. Temporal logic based models/systems

The essential role of time in the modeling of natural processes has given rise in recent years to a body of artificial intelligence research into temporal theory. This research has led to a variety of temporal systems, attempting to capture the primary elements of time. However, as time goes on, the world may change its state from one into another, triggered by some certain events or processes that take place over time. Although different temporal systems show considerable commonality in structure, they also show considerable differences in formalization. In the literature, there are three choices regarding the primitive for the ontology of time: instantaneous points, durative intervals and both points and intervals and problems may arise when one conflates different views of temporal structure. A natural approach to representing and reasoning about the actions, events, processes is to associate them with time elements (i.e., instantaneous points and/or durative intervals) [18].

Many theories, [3], [22] and [23], are based on points as the basic primitive element. In these theories, intervals are defined in terms of points, usually by means of beginning and ending points. However, as Allen has commented [7], modeling intervals by taking their bounding-points can lead to problems: the annoying question of whether bounding-points are in the interval or not must be addressed, seemingly without any satisfactory solution. If intervals are all closed then adjacent intervals have bounding-points in common, which when adjacent intervals correspond to states of truth and falsehood of some property, can lead to situations in which a property is both true and false at an instant. Similarly, if intervals are all open, there will be points at which the truth or falsity of a property will be undefined. The solution, in which intervals are all taken as semi-open, so that they sit conveniently next to one another, seems arbitrary and unsatisfactory. Other theories, predominantly in [7], [8], treat intervals as primitive, and in [13], [14], treat both intervals and points as primitive on an equal footing.

After carefully considering the literature, we found that formal grounding is absent in the main commercial modeling techniques. However, frameworks based on temporal logic are available in the literature that are expressive than others and provides formal semantics if mapped. The objective of our work is to provide an ontology mapping of BPMN, UML AD and EPC to temporal model for business processes [18] which provides formal semantics.

III. MAIN BPM TECHNIQUES AND THEIR MODELING TERMS/CONSTRUCTS

Before describing any technique we define what a business process is. According to Davenport [24], processes are defined as “structured, measured sets of activities designed to produce a specified output for a particular customer or market”. There are so many other definitions but in essence all are the same: processes are relationships between inputs and outputs, where

⁵ IBM, SAP AG, *WS-BPEL Extension for People–BPEL4People*. Whitepaper, 2005

[<http://www.ibm.com/developerworks/webservices/library/specification/wsbpe4people/>].

⁶ WfMC. Process Definition Interface – XML Process Definition Language, version 2.00, October 2005. [http://www.wfmc.org/standards/docs/TC-1025_xpdl_2_2005-10-3.pdf]

inputs are transformed into outputs using a series of activities, which add value to the outputs.

It is beyond the scope of this paper to go into further discussion on the difference between business processes and processes in general. It seems that some authors take them as synonymous. For example, in contrast to [15], [16] defines business process as “a collection of activities that takes one or more kinds of input and creates an output that is of value to the customer”.

A. The Main Modeling Terms/Constructs of BPMN

The main goal of BPMN is to standardize a business process modeling notation in order to provide a simple means of communicating process information among business users, customers, suppliers, and process implementers. The basic BPMN constructs are *activity*, *event*, *gateway* and *sequence flow*.

An *activity* is a generic term for work performed within a company. It can be atomic or non-atomic (compound). The types of activity are: *task* and *process/sub-process*. A *task* is an atomic activity included within a *process/sub-process*. *Process/Sub-Process* is a sequence or flow of activities in an organization with the objective of carrying out work. There are two basic types of *processes*; *Private (internal)*, and *Public Processes*. The *Private Processes* are those internal to a specific organization and further divided into two types *Executable* and *Non-executable*. An *executable process* is modeled for the purpose of being executed according to the semantics defined and there will be stages in lifecycle of a process where not enough detail available to execute it. A *nonexecutable process* is modeled for the purpose of documenting process behavior at a modeler-defined level of detail. A *public process* represents the interactions between a private business process and another process or participant. In addition to *process*, there are two types of *sub-processes*: *embedded* and *independent*. The *embedded sub-processes* are further divided into two views of *sub-processes*; *collapsed* and *expanded*. *Collapsed* view hides its details or *expanded* view that shows its details within the view of the *process* in which it is contained. An *event* is something that “happens”, like a trigger or a result, during the execution of a business process affecting the flow of the process. Since an event can start, suspend, or end the flow, we can distinguish the above between *start events*, *intermediate events*, and *end events* respectively. A *gateway* is a modeling element used to represent the interaction of different sequence flows, as they diverge and converge within a process. When the sequence flows arrive at a *gateway*, they can be merged together on input and/or split apart on output. There are different types of *gateway* according to the types of behavior they define in the sequence flow. Decisions and branching are represented by *OR-Split*, *exclusive-XOR*, *inclusive-OR*, and *complex*, merging is represented by the *OR-Join gateway*, and forking is represented by the *AND-Split gateway*, and joining by the *AND-Join gateway*.

However, there are other constructs used to model any relevant entity that is able to activate or perform a process i.e. *pool* and *lane*. They are representing more aggregate organization units and more specific ones, respectively. They

allow for a partitioning of activities according to the performers.

Critical Evaluation BPMN: BPMN elements are hard to sketch on paper unlike UML AD or flowcharts [21]. In [4], numerous ambiguities in the descriptions and under specifications of semantically relevant concepts pervade the standard document and leave space for incompatible (but, due to the lack of precision, standard ‘conforming’) interpretations in design, analysis and use of BPs. Another under specification concerns *expression evaluation*. When should expressions (particularly event expressions) be evaluated? Depending on the type, it could (probably should) be either before or at process start, or upon state change or when a token becomes available. BPMN provides only poor conceptual support for numerous features which are characteristic of the design and management of business processes. One of BPMN's main shortcomings is that an increase in graphical notation yields an increased complexity in the meta-model and makes transparent faithful implementation more and more impractical. BPMN comes with a plethora of interdefinable constructs. Instead of defining a core of independent constructs in terms of which other constructs can be defined, as suggested in [5] for a previous version of the standard (it was also suggested to the standardization committee). The fuzzy overlapping of different constructs prevents ‘closed’ descriptions of individual constructs in one place and makes their comprehension unnecessarily complex by forcing the reader to simultaneously and repeatedly consider multiple sections of the standard document. It also creates the problem that where the definitions overlap they have to be consistent; this problem is not considered in the standard document. Furthermore a statistical evaluation (of BPMN 1.1) shows that ‘the average BPMN model uses less than 20% of the available vocabulary’ and that, out of the more than 50 graphical elements in BPMN, ‘Only five elements (normal flow, task, end event, start event, and pool) were used in more than 50% of the models we analyzed.

B. The Main Modeling Terms/Constructs of UML (AD)

UML(AD) is the object-oriented equivalent of flow charts and data-flow diagrams from structured development and describes the workflow behavior of a system. The process flows in the system are captured in the activity diagram and also illustrates the dynamic nature of a system by modeling the flow of control from activity to activity. The main constructs of UML AD are *activity*, *action*, *initial node*, *final activity node*, *connecting nodes*, and *control flow*.

An *activity* is used to represent a set of actions and an *action* represents a single step within an activity. An *Initial Node* is the entry point to an activity diagram and an *Activity Final Node* is the final node in an activity that terminates the actions in that activity. However, an *Object Node* is used to represent an object that is connected to a set of Object Flows. A *Decision Node* is used to represent a test condition to ensure that the control flow or object flow only goes down one path. A *Merge Node* is used to bring back together different decision paths that were created using a decision-node. A *Fork Node* is used to split behavior into a set of parallel or concurrent flows of activities (or actions). A *Join Node* is used to bring back together a set of parallel or concurrent flows of activities (or

actions). *Control flow* shows the flow of control from one action to the next and also shows the sequence of execution.

Critical Evaluation of UML AD: Weaknesses of UML AD are given below:

- Some of the UML AD constructs lack a precise syntax and semantics. For instance, the “well-formedness” rules linking forks with joints are not fully defined, nor are the concepts of dynamic invocation and deferred events, among others.
- UML ADs are extremely limited in modeling resource-related or organizational aspects of business processes. It is interesting to note that UML ADs cannot capture many of the natural constructs encountered in business processes such as cases and the notion of interaction with the operational environment in which the process functions.

These limitations observed by [19] and [20] are common to many other business process modeling formalisms and reflect the overwhelming emphasis that has been placed on the control-flow and data perspectives in contemporary modeling notations. While UML ADs are functional, business analysts somehow cannot use them without prior technical knowledge in [1]. A business analyst cannot model a business process and its sub-processes from the highest level to the lowest level of detail in an UML AD. It is increasingly losing favor with practitioners (although there are currently several projects working on UML-to-BPEL translations by IBM and OMG) [17]. This is mainly due to industry’s growing consolidation of BPMN as the de facto standard for BPM.

C. The Main Modeling Terms/Constructs of EPC

Event-driven process chains (EPCs) are part of a holistic modeling approach, called the ARIS framework. Process modeling uses event-driven process chains. The main building blocks of event-driven process chains are *events, functions, connectors, and control flow edges*.

Functions describe transformations from an initial state to a resulting state. They represent units of work and granularity of these functions depends on the modeling purpose. The entering of a business relevant state is represented by an *event*. *Events* trigger *functions* and passive elements in EPC. They describe under what circumstances a *function* works or which state a *function* results in. Examples of *events* are "requirement captured", "material in stock", etc. EPC diagram must start with an *event* and end with an *event*. Unlike events, functions are active elements that take input and transform it to output. Functions can also make decisions that influence the behavior of the process through connector nodes associated with the function. They are triggered by events, and on the completion of a function, an event occurs. There are three kinds of logical relationships exists between events and functions and are *branch/merge, fork/join* and *or*. *Control flow* connects events with functions, process paths, or logical connectors creating chronological sequence and logical interdependencies between them.

Critical evaluation of EPC: It works as an ordered graph of events and functions and supports parallel execution of

processes. However, the semantics and syntax of the EPC are apparently not well defined [6] [26]. Because of these limitations and the absence of a standardization process, the EPC will not be classified as a graphical standard.

Formal systems are used for the clarity and unambiguous reasoning and representation. However the commercial modeling tools and techniques are ambiguous in representation and lack formal foundation. To provide a clear and precise meaning to different modeling terms discussed above, will provide a mapping of ontology to formal logic. For this, we categorize and group modeling terms/constructs used by the modeling tools and techniques discussed above into three distinguished conceptual terms and formally defining them. This categorization will serve the purpose of the ontology mapping to a formal system.

D. Conceptual Categorization of Modeling Terms/Constructs

We have seen that the modeling terms discussed in previous sub-sections have somewhat similarities and differences. Main modeling techniques such as BPMN, UML AD and EPC provide no formal foundation and leads to ambiguity in the design. To overcome this problem and fill this gap, we propose conceptual categorizations of modeling terms used by BPMN, UML AD and EPC. Modeling terms of these techniques are grouped into three conceptual categories: *process, connector* and *constraint*. These categories refer to the modeling terms/constructs used in modeling the dynamic and static aspects of the domain. This categorization will assist in ontology mapping of commercial modeling terms to formal notion given in section V. We will take conceptual category’s first letter of every word to form a title of our categorizations which would be *Process, Connector* and *Constraint (PCC)* and formally defined below:

Definition (Process): In our proposed conceptual category of *process*, we take every activity as a non-instantaneous *process* and define *process* as nonempty set containing processes, or it also could be a singleton set which may be considered as a special case of a *process* and above can be represented as

$$\text{Process 'P'} \neq \emptyset \text{ or Process 'P'} = \{p\}$$

In addition, an event may also be considered as a *special process* that starts (p_s) and ends (p_e) a *process* P and is shown below:

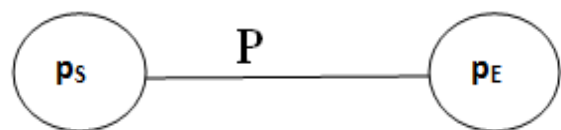


Fig. 1. An abstract process

Keeping in mind the temporal nature of the *process*, we could formalize it by using a predicate **Occurring**, for *process* P i.e. non empty set of processes, occurring over a time element *T* and using **Dur** to express the duration by the following axiom:

Occurring $(P, T) \Rightarrow \exists p_S, p_E \wedge \text{Dur}(p_S) = 0 \wedge \text{Dur}(p_E) = 0 \wedge$
 Meets $(p_S, T) \wedge \text{Meets}(T, p_E) \wedge \text{Occurring}(p_S, P) \wedge \text{Occurring}(P, p_E)$(Axiom 1)

p_S and p_E are the *special processes* i.e. events, that starts and ends a process P respectively. The above axiom refers to a general process which occurs over a time element that may be divisible.

To formalize a singleton set of process p , using **Occurring** predicate, **Dur** for duration and occurring over a time element t is given below in axiom 2 using temporal relation ‘In’ given by Allen in [8]:

Occurring $(p, t) \Rightarrow \exists p_S, p_E \wedge \text{Dur}(p_S) = 0 \wedge \text{Dur}(p_E) = 0 \wedge$
 Meets $(p_S, t) \wedge \text{Meets}(t, p_E) \wedge \neg \exists [t_1 \wedge \text{Dur}(t_1) > 0 \wedge \text{In}(t_1, t) \wedge$
 Occurring $(p, t_1)]$(Axiom 2)

Axiom 2 has shown a singleton i.e. nondivisible, process. *Special process* i.e. event, may also occur when finishing a process and starting a new process i.e. start, end and intermediate event of BPMN. We have shown above that our conceptual category *Process* is general enough to map BPMN’s modeling terms/constructs *task, process/sub-process, start event, intermediate event* and *end event*, UML AD’s *activity, action, initial node* and *final activity node*, and EPC’s *functions* and *events*.

Definition (Connector): We describe Connector as a set which contains logical operators AND and OR, and is given below:

$$\text{Connector 'C'} = \{\wedge, \vee\}$$

Our proposed conceptual category *connector* can map BPMN’s modeling term/construct *gateway*, UML AD’s *connecting nodes*, and EPC’s *logical connectors*.

Definition (Constraint): We describe Constraint as a set 30 derived relation given in [13] gathered in four groups and is given below:

Constraint ‘C’ = {point to point, point to interval, interval to interval, interval to point}

The conceptual category *constraint* can map BPMN’s modeling term/construct *sequence flow*, UML AD’s *control flow* and EPC’s *control flow*.

The formalization of three conceptual categories has paved the way to group main modeling terms of commercial modeling techniques. This effort provides a generalized view of all modeling terms of the modeling techniques discussed in this paper. Now, we would show the aforementioned **PCC** categorization and its subsumption of modeling terms in the table I.

IV. TEMPORAL MODEL FOR BUSINESS PROCESSES

Ma and Knight [13] have proposed more general time theory that considers point and interval both on equal footing i.e. primitives. In addition, abstract modeling terms have been proposed [18] and we will use them for ontology mapping purposes. BPM terms such as *action, event, business process, sub-process* and *temporal relations* are defined by providing formalisms and discussed in the next sub-section.

TABLE I. CONCEPTUAL CATEGORIZATION OF BUSINESS PROCESS MODELING TERMS

| Modeling Notation | Modeling Category | Modeling Terms |
|-------------------|-------------------|---|
| BPMN | Process | task, process/sub-process, start event, intermediate event, end event |
| | Connector | gateways: AND, OR & XOR |
| | Constraint | sequence flow |
| UML AD | Process | activity, action, initial node, final activity node |
| | Connector | connecting nodes: Merge Node, Fork Node, Join Node |
| | Constraint | control flow |
| EPC | Process | function, event |
| | Connector | logical connectors: AND, OR, XOR |
| | Constraint | control flow |

Conceptual Categorization

A. Abstract Business Modeling Terms

BPM terms *action, process, and sub-processes* [18] are associated with non-instantaneous activity and *event* is associated with instantaneous activity i.e. point. An *action name* is an identifier that describes a certain type of non-instantaneous activity. For instance, “push a cart”, “cut wire” and so on. They used a, a_1, a_2, \dots , etc., to denote action names, and write the set of action names as A . Without confusion, they simply call an action name, say a , action a . It is important to note that a given type of action may perform once, more than once over different time moments, or may not even perform at all. An *event name* is an identifier that describes a certain type of instantaneous activity. For instance, “departure at”, “start cut wire”, “finish cut wire” and so on. They used e, e_1, e_2, \dots , etc., to denote event names, and write the set of event names as E . Without confusion, we may simply call an event name, say e , event e . Like action, a given type of event may occur once, more than once at different time points, or may not even occur at all. A *business process name* is a set of action names and set of event names.

They have used logical connectors to establish the relation between action and event i.e. \wedge, \vee . Allen and Hayes [9] introduced 13 relations in his famous Interval Algebra; however Ma and Knight [13] from these 13 relations derived 30 temporal relations to constrain the order of the actions and events. In terms of the single primitive relation *Meets*, other binary relations over points/intervals can be classified into 4 groups:

- Point – Point: {Equal, Before, After}
- Point – Interval: {Before, After, Meets, Met_by, Starts, During, Finishes}
- Interval – Point: {Before, After, Meets, Met_by, Started_by, Contains, Finished_by}

- Interval – Interval: {Equal, Before, After, Meets, Met_by, Overlaps, Overlapped_by, Starts, Started_by, During, Contains, Finishes, Finished_by}

The next section will provide ontology mapping based on the categorization provided in section III.

V. ONTOLOGY MAPPING

To provide formal semantics for commercial modeling techniques and languages mentioned in section III, there are different extensions provided separately in the literature. However, there is no effort has been made to provide an ontology mapping for them.

There are different framework/systems available in the literature which discusses the ontology of modeling tools and techniques i.e. process, event and action. Our effort is to achieve a comprehensive ontology mapping using **PCC** categorization introduced in section III that will map commercial modeling terms to formal modeling terms in [18] and shown in table II.

TABLE II. ONTOLOGY MAPPING TABLE FOR BPM TERMS/CONSTRUCTS

| Modeling Notation | Modeling Category | Modeling Terms | Abstract Modeling Terms |
|-------------------|-------------------|---|-------------------------------------|
| BPMN | Process | task, process/sub-process, start event, intermediate event, end event | action, event, process |
| | Connector | gateways: AND, OR & XOR | logical operators \wedge , \vee |
| | Constraint | sequence flow | temporal relations |
| UML AD | Process | activity, action, initial node, final activity node | action, event, process |
| | Connector | connecting nodes: Merge Node, Fork Node, Join Node | logical operators \wedge , \vee |
| | Constraint | control flow | temporal relations |
| EPC | Process | function, event | action, event, process |
| | Connector | logical connectors: AND, OR, XOR | logical operators \wedge , \vee |
| | Constraint | control flow | temporal relations |

Ontology Mapping

The above table can also be shown in fig 2:

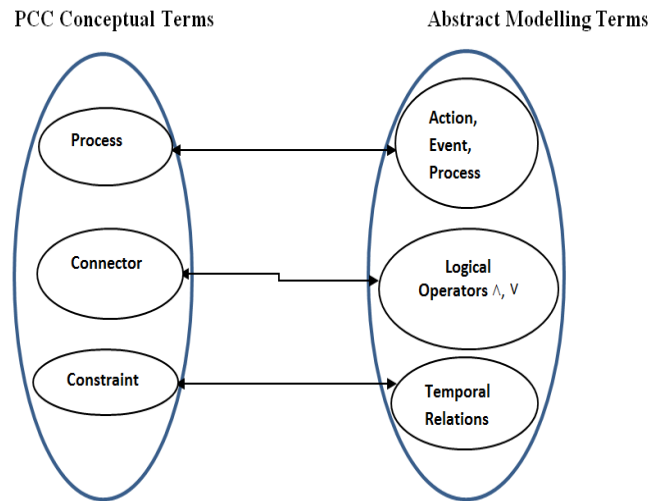


Fig. 2. Ontology Mapping

Ontology mapping using **PCC** categorization provides a platform for business process modeling techniques to have formal semantics which are intuitive but formal. This attempt will lead to provide a step towards foundation for business process modeling based on temporal logic.

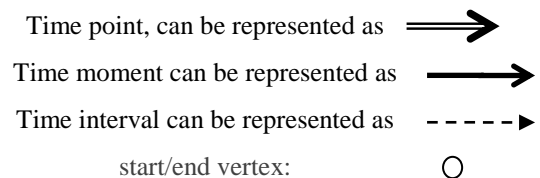
To show the mapping is comprehensively achieved the aim of this paper; we will discuss an illustrative example in next section graphically representing it using modeling techniques used in section III and IV.

VI. GRAPHICAL REPRESENTATION

The time theory outlined above in terms of the single *Meets* relation allows a simple graphical representation of any set of temporal knowledge. Each time element t is denoted as a directed arc of the graph labeled by t (and its duration if it is known), with a pair of nodes which are called the start-vertex, and the end-vertex, of the arc, respectively.

- Each relation $Meets(t_i, t_j)$ is represented by means of merging the end-vertex of t_i and the start-vertex of t_j as a common vertex, of which t_i is an in-arc and t_j is an out-arc, respectively. In this case, arc t_i is said to be adjacent to arc t_j .

In general, the temporal order relation between two time elements may be given in any form of those 30 as classified in section IV. However, as defined above in section IV, each of these temporal relations can be derived from the single *Meets* relation. Therefore, all the knowledge about the temporal relations over a given collection of time elements (points and/or intervals) can be transformed and stored as a table of *Meets* relations in the knowledge base. We use the following graphical representation to represent the available knowledge:



This graphical representation of the underlying logical structure forms the link between the temporal theory, and practical business process diagrams. For instance, will consider a process graphically represented in the aforementioned three commercial BPM languages and afterwards will represent the same process graphically using logically defined terms to show the expressiveness, simplicity and generality of the logical structure.

A. An Illustrative Example:

BPMN: Fig 3 describes an example business process using BPMN. The business process begins with a start event to execute the first task **Record the Claim**. **Calculate the Insurance Sum** and **Record the Claim** are part of the Pool **Financial Claim Specialist**. After the task **Record the Claim** the pool **Claim Administrator** is responsible for the process. The exclusive gateway splits the flow, because the decision has to be made if the insurance sum has a minor amount or major amount. If the insurance sum has a minor amount, then only task **Contacting the Garage** is processed. But if the insurance sum has a major amount then **Contacting the Garage** concurrently starts with the **Checking History of the Customer**. An inclusive gateway combines the different paths. The gateway conforms to the logical operator OR. After the task **Examination of Results** the decision has to be made if the company **Pay for the Damage** or **Does Not Pay for the Damage**. After that decision the case is closed.

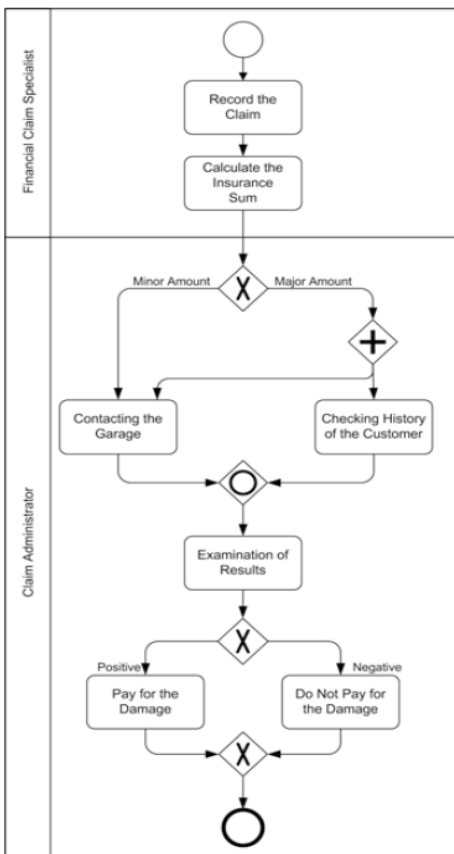


Fig. 3. Business Process using BPMN

UML AD: A business process example in UML AD is shown in fig 4. The business process starts with an *initial node*, to activate the first action; **Record the claim**, Record the claim passes the token to the next action to **Calculate the Insurance Sum**. These two actions part of activity partition **Financial Claim Specialist**. After calculating the insurance sum the path is split up by a decision node into two alternative flows, depending if the insurance sum has a minor amount then the action **Contacting the Garage** starts. If the insurance sum has a major amount, then the flow is split up into parallel paths by a fork node. That means that the actions **Contacting the Garage** and **Checking History of the Customer** are executed concurrently. A merge node combines the different flows, and accepts the token as well as of one path or both paths. The action **Examination of Results** decides that the claim is handled either positive or negative. Therefore a decision node splits up the path in two alternative flows, with the actions **Pay for Damage** or **Do Not Pay for Damage**. After that decision the business process ends with a flow final activity node.

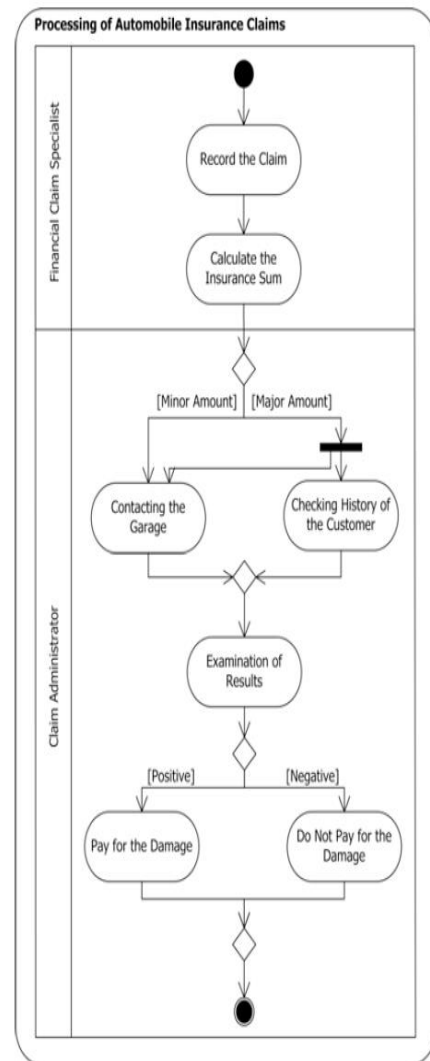


Fig. 4. Business Process example using UML AD

EPC: Fig 5 shows a business process example in EPC. The business process starts with event **New Claim Submitted**. The function **Record the Claim** starts after the first event. After the event **Claim Recorded** the function **Calculate the Insurance Sum** begins. The organizational role **Financial Expert** is responsible for the functions **Record the Claim** and **Calculate the Insurance Sum**, the path of business process splits up into alternative flows, depending of the insurance sum has a **Minor Amount** or **Major Amount**. If the insurance sum has a minor amount, then the function **Contacting the Garage** starts. If the insurance sum has major amount then the functions, **Contacting the Garage** and **Checking History of the Customer** starts concurrently. These two functions are connected with the next event. **Results Collected** by an OR-Join. At that time the organizational role **Claim Administrator** is responsible for the business process. Due to the fact that either **Checking History of the Customer** or **Contacting the Garage** is executed, or both functions are processed, the OR-Join is needed. If the results are collected, then the function **Examination of Results** starts processing, to decide if the claim is handled **Positive** or **Negative**. If the claim is handled positive, then the insurance company **Pays for the Damage**, otherwise not. In both situations **Case is Closed**.

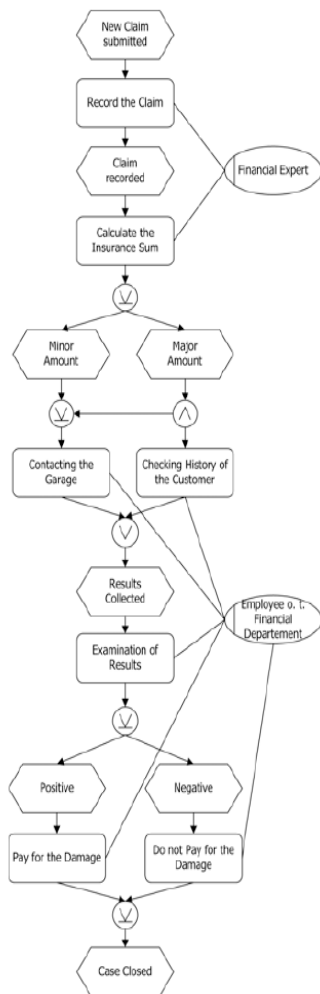


Fig. 5. Example business process of an EPC

After using **PCC** categorization and graphical representation described above, the aforementioned example can be shown in fig 6, using abstract modeling terms/constructs. Notice that there are both absolute and relative times in this model showing temporal ordering of processes.

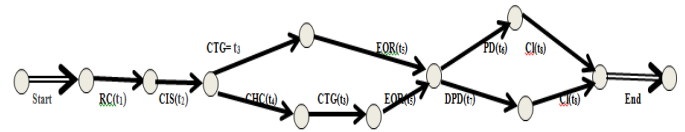


Fig. 6. Business Process using Abstract Modeling constructs

Note: RC: Record the Claim: t_1 ; CIS: Calculate the Insurance Sum: t_2 ; CTG: Contacting the Garage: t_3 ; CHC: Checking History of the Customer: t_4 ; EOR: Examination of Results: t_5 ; PD: Pays for the Damage: t_6 ; DPD: Does Not Pay for the Damage: t_7 ; CI: Closed the Case: t_8 .

Using *Meets* relation, the above can be represented as

$$\text{Meets}(t_1, t_2) \wedge (\text{Meets}(t_2, t_3) \wedge \text{Meets}(t_3, t_5) \wedge \text{Meets}(t_5, t_6) \wedge \text{Meets}(t_6, t_8)) \vee (\text{Meets}(t_2, t_4) \wedge \text{Meets}(t_4, t_3) \wedge \text{Meets}(t_3, t_5) \wedge \text{Meets}(t_5, t_7) \wedge \text{Meets}(t_7, t_8))$$

We have seen above that abstract modeling terms have subsumed all modeling elements of aforementioned commercial modeling techniques shown in fig 3, 4 and 5. We also have used a simpler graphical representation to verify the logical structure that is intuitive and expressive than others to represent concepts and knowledge in a simpler but composed way.

VII. CONCLUSION & FUTURE WORK

In this paper we presented a framework for BPM using ontological mapping. A versatile conceptual categorization introduced and subsequently formalized the categories that can group and map modeling terms used by different tools and techniques to formal notion. There are a very large number of BPM languages so our effort is to work towards bridging the gap for business process modeling to be grounded on formal logic. Some efforts carried out in providing formal semantics and have given partial results so far. We have consolidated it by proceeding along two lines. One is methodological, necessary to provide its ontological mapping, and another is empirical, necessary to check its value in real business settings as given below:

- the fact that BPMN, UML AD and EPC does not provide formal grounding to represent their constructs, being it specified in an informal way;
- Ontology mapping for business process modeling based on logic provides a step towards in providing a formal foundation.

The absence of formal grounding in modeling techniques will always result in some loss of data or semantics of the control flow. After a careful analysis of the literature we identified temporal logic as a possible candidate to match the above requirements. To fill this gap, it is envisaged by the author to provide an axiomatic system based on temporal logic

for the business process modeling that will formally ground the modeling notion and provide a unified graphical representation for process which would be simpler and easy to design and more suited to majority of the user's needs.

ACKNOWLEDGMENT

We are thankful to Vice Chancellor of University of Greenwich who has sponsored this project.

REFERENCES

- [1] A.E. Bell. Death by UML fever: self-diagnosis and early treatment are crucial in the fight against UML fever", ACM Queue, Vol. 2(1), 2004, pp. 72-80.
- [2] A.-W. Scheer, O. Thomas and O. Adam. *Process Modeling Using Event-Driven Process Chains*. In "Process-Aware Information Systems", edited by M. Dumas, W. hal-00656686 van der Aalst, A.H.M ter Hofstede. Wiley-InterScience, 2005, pp 119-145.
- [3] B. C. Bruce. A Model for Temporal References and Application in a Question Answering Program. *Artificial Intelligence*, An International Journal Vol 3, 1972, pp.1-25.
- [4] E. Börger. Approaches to modeling business processes: a critical analysis of BPMN, workflow patterns and YAWL." *Software & Systems Modeling* Vol 11(3), 2012, pp 305-318.
- [5] E. Börger and B. Thalheim. A method for verifiable and validatable business process modeling. In *Advances in Software Engineering*, Vol 5316 of LNCS, 2008, pp 59-115. Springer-Verlag.
- [6] E. Kindler. On the semantics of EPCs: a framework for resolving the vicious circle, paper presented at Business Process Management: 2nd International Conference, 2004, BPM 2004, Potsdam.
- [7] J. Allen. Maintaining Knowledge about Temporal Intervals. *Communication of ACM*, Vol 26, 1983, pp.123-154.
- [8] J. Allen. Towards a General Theory of Action and Time, *Artificial Intelligence*, Vol 23, 1984, pp 123-154.
- [9] J. Allen, and J. Hayes. Moments and Points in an Interval-based Temporal-based Logic, *Computational Intelligence*, Vol 5(4), 1989, pp 225-238.
- [10] J. Harrington. *Business Process Improvement: The Breakthrough Strategy for Total Quality, Productivity and Competitiveness*. McGrawHill, New York, USA, 1991.
- [11] J. L. Peterson. Petri Nets. *ACM Computing Surveys* Vol 9(3):1977, pp 223-252.
- [12] J. L. Peterson. *Petri Net Theory and the Modeling of Systems*. Prentice Hall, 1981.
- [13] J. Ma and B. Knight. A General Temporal Theory, *The Computer Journal*, Vol 37(2), 1994, pp 114-123.
- [14] M. B. Vilain. A System for Reasoning about Time. *Proceedings of AAAI*, Vol 1, 1982, pp.197-201.
- [15] M. Hammer. Reengineering work: Don't automate. Obliterate. *Harvard Business Review* Vol 68 (4), 1990, pp 104-112.
- [16] M. Hammer and J. Champy. *Reengineering the Corporation: A Manifesto for Business Revolution*. NewYork, USA, 1993
- [17] M. Koskela, and J. Haajanen. Business process modeling and execution: tools and technologies report for the SOAMeS project", VTT Research Notes No. 2407, VTT Technical Research Centre of Finland, Espoo, 2007
- [18] .M. Petridis, J. Ma, and B. Knight. Temporal model for business Process" *Intelligent Decision Technologies*, Vol 5(4), 2011, pp 321-331.
- [19] N. Russell, W.M.P. Van Der Aalst, A.H.M. Ter Hofstede, and P. Wohed. On the suitability of UML 2.0 activity diagrams for business process modeling, *Proceedings of the 3rd Asia-Pacific Conference on Conceptual Modeling*, Australian Computer Society Vol. 53, 2006, pp. 95-104.
- [20] P. Wohed. *Pattern-based Analysis of UML Activity Diagrams*, Beta, Research School for Operations Management and Logistics, Eindhoven, 2004.
- [21] P. Wohed, W. M. P. van der Aalst, M. Dumas, A.H.M. ter Hofstede, and N. Russell. On the Suitability of BPMN for Business Process Modeling, *Business Process Management*, Vienna, 2006, pp. 161-76.
- [22] R. Dechter, I. Meiri and J. Pearl. Temporal Constraint Networks. *Artificial Intelligence*, Vol 49, 1991, pp.61-95.
- [23] R. Maiocchi. Automatic Deduction of Temporal Information. *ACM Transactions on Database Systems*, Vol 4, 1992, pp.647-688.
- [24] T.H. Davenport. *Process Innovation: Reengineering Work through Information Technology*. Harvard Business School Press, Boston, MA, USA, 1993.
- [25] M. Weske, *Concepts, Languages, Architectures*. Vol. 14. 2007, Berlin: Springer-Verlag.
- [26] W.M.P. Van der Aalst. Formalization and verification of event-driven process chains", *Information and Software Technology*, Vol 41(10), 1999, pp. 639-50.

Adaptive Cache Replacement: A Novel Approach

Sherif Elfayoumy
School of Computing
University of North Florida
Jacksonville, Florida

Sean Warden
School of Computing
University of North Florida
Jacksonville, Florida

Abstract—Cache replacement policies are developed to help insure optimal use of limited resources. Varieties of such algorithms exist with relatively few that dynamically adapt to traffic patterns. Algorithms that are tunable typically utilize off-line training mechanisms or trial-and-error to determine optimal characteristics.

Utilizing multiple algorithms to establish an efficient replacement policy that dynamically adapts to changes in traffic load and access patterns is a novel option that is introduced in this article. A simulation of this approach utilizing two existing, simple, and effective policies; namely, LRU and LFU was studied to assess the potential of the adaptive policy. This policy is compared and contrasted to other cache replacement policies utilizing public traffic samples mentioned in the literature as well as a synthetic model created from existing samples. Simulation results suggest that the adaptive cache replacement policy is beneficial, primarily in smaller cache sizes.

Keywords—cache replacement policy; high performance computing; adaptive caching; Web caching

I. INTRODUCTION

Caching in computing has been a proven form of performance enhancement for some time, most notably in memory paging [1] [2]. The basic premise is that objects that are frequently used or most likely to be used can be stored in a location that provides performance benefits by virtue of being temporally or locally near the consumer of the object. Temporally meaning that an object may be served more quickly from a given location when compared to the service time from another host location, possibly due to reduced latency of access from faster resources (disk, memory, etc.) or increased service bandwidth. Because cache resources are finite in size, however, one problem in caching is that there exists a wide variety of algorithms for replacing objects in a filled cache.

Effective caching algorithms are necessary to insure users experience favorable performance benefits, especially in an environment as diverse and distributed as the Internet. Such performance benefits are a reduction in the delay between the time a user requests an object and its delivery to the user.

Since the Web has evolved, several caching algorithms have been suggested in literature; Balamash and Krunz's survey of replacement algorithms described no fewer than twenty different replacement schemes [3]. Each strategy utilizes different parameters to determine objects to replace – Balamash classified caching policies based on whether they utilized frequency, recency, or size information. While some algorithms utilize a single traffic trait, others employ a

functional approach to compute a derived 'cost' of an object cache miss based on multiple parameters, thus removing the lowest 'cost' object when a replacement is required. Not surprisingly, the research has shown that functional approaches are generally more computationally complex than those based on a single attribute [3]. Due to the variety of approaches, metrics, and parameters utilized (sometimes singly and sometimes in combination), each algorithm has distinct performance characteristics.

Since access patterns are unique to each environment, certain algorithms are more suitable than others depending on the traffic situation, and one single algorithm is not best in class for all situations. A web server that functions as a search engine will have different traffic patterns, and thus unique cache architecture requirements, when compared to a university web portal, online encyclopedia, multimedia server, etc. [4]. Algorithm designers (such as [5]) have attempted to overcome this problem using one or more static parameters that can be tuned offline to optimize performance through analysis of historical object requests.

Web caches are designed to provide apparent speed benefits to object requestors by offloading objects from the web server itself to a location that is physically or logically closer to the requestor and/or decreases the amount of load experienced by the web server [6]. A web cache typically stores its objects in some form of memory or disk. Because these storage resources are finite, however, cache replacement policy algorithms are utilized to determine which objects to remove from the cache as new objects are accessed which are deemed more productive to cache [7]. Ideally, these replacement policy algorithms will always choose to keep the objects that will provide the best performance.

Developing an algorithm that is optimal for all traffic patterns is a challenging problem. Some algorithms attempt to overcome this problem using static tuning parameters that are set using offline training, trial-and-error, or possibly via an educated guess. Others attempt to adapt dynamically, using computationally complex dynamic parameters that are based on historical object requests. Because of the ubiquity of caching across a variety of computing environments – microprocessors, web, thin-clients, wireless devices, etc. – and the wide variety of possible differences in traffic and access patterns within and across these environments, researching methods for adaptive performance optimization is a worthy objective.

In this article, a novel adaptive approach to finding an efficient cache replacement policy for a given traffic pattern by sectioning the cache storage space into areas managed by

separate object replacement policies is introduced and assessed. The overall cache space then utilizes a tuning algorithm to allow the overall cache to choose the best policy based on current access patterns. This effort focuses on caching within a web environment. However, the approach is extensible to other caching environments. In section II an overview of the most prominent cache replacement policies is provided, and in section III the adaptive cache replacement policy is described. The assessment approach and preliminary results are described in section IV. Last, section V highlights the conclusions.

II. CACHE REPLACEMENT POLICIES

A variety of cache replacement policy algorithms have been designed and evaluated in literature, with a goal of maximizing cache effectiveness as measured using performance metrics such as Hit Ratio. Each algorithm generally utilizes one or more of three pieces of information about requested objects: recency of access, frequency of access, or size. Algorithms range from those which are very simple in that they only use a single traffic parameter, such as LRU and LFU, to those which are very complex, using multiple parameters as well as statically tuned constants, such as Hybrid and GDSF. In the case of LRV, dynamic probability functions – built based on static algorithm analysis – are included but at the cost of implementation challenges and computational complexity.

Evaluating the performance of replacement algorithms is typically accomplished using real-world web logs (also known as web traces). These web logs are cleansed to remove non-cacheable requests (such as cgi or other dynamically generated content), then run through a program to simulate the replacement algorithm by ‘playing back’ the cleansed web log. The simulation process calculates benchmarks over a range of cache sizes. The most common metrics are hit ratio (HR), or the ratio of cache hits to all requests; byte hit ratio (BHR), or the ratio of bytes returned from the cache to all bytes requested; and latency ratio (LR), or the delay experienced by a user for objects retrieved from the cache verses that experienced if no objects were cached. Other effectiveness measures include reduced packets, or the ratio of network packets avoided due to caching to the total packets that would have otherwise been seen by the server; and reduced hops, a similar measure that focuses on network hops between client and server. Balamash and Krunz define the most common measures as follows [3]:

$$HR = \frac{\sum_{i \in R} h_i}{\sum_{i \in R} f_i} \quad BHR = \frac{\sum_{i \in R} s_i \cdot h_i}{\sum_{i \in R} s_i \cdot f_i} \quad LR = \frac{\sum_{i \in R} d_i \cdot h_i}{\sum_{i \in R} d_i \cdot f_i}$$

Where: s_i = size of object i ; f_i = total number of object requests for object i ; h_i = total number of cache hits for object i ; d_i = average server retrieval delay for object i ; R = set of all requested objects.

A. Least Recently Used (LRU) Algorithm

The Least Recently Used algorithm essentially uses a single parameter to decide which object to remove from the

cache: time since last access. The basic premise of the algorithm is that those objects that are most likely to be accessed will have been accessed more recently than those that are not as likely. While simple to implement and requiring less computational power than most other algorithms, LRU has been outclassed by several other replacement algorithms: Balamash and Krunz’s experiments showed that for large cache sizes, LUV, GDS, and Hyper-G produced better results for both HR and LR [3], while Bahn et al. found that for large cache sizes, LUV, Hybrid, Size, Mix, and sw-LFU performed better for HR and LUV was better for LR [5].

B. Least Frequently Used (LFU) Algorithm

Another relatively simple algorithm, Least Frequently Used utilizes a frequency counter for each object in the cache. Objects that are most frequently accessed are thus more likely to remain in the cache and presumably provide benefit to future users. Similar to LRU in terms of ease of implementation and complexity, this algorithm has also been surpassed by algorithms that are more efficient; Bahn et al. noted that LUV, LNC-R-W3, GDS, and LRV outperformed LFU for HR, BHR, and LR [5].

C. Size Algorithm

The Size algorithm is another simple algorithm - size is the only measure utilized for eviction evaluation [8]. When an eviction is required, Size removes the largest object currently in the cache with the idea that users are less likely to re-request larger objects. Smaller objects, then, are more likely to remain in the cache long-term [3]. Additionally, this allows for a larger number of objects to remain in the cache, potentially improving hit rates for some traffic patterns. Algorithm implementation is simple when compared to algorithms using multiple parameters, but exhibits generally poor performance: Balamash et al. found that the Size algorithm was a middle-of-the-pack performer for HR and absolute worst for BHR and LR using a simulated DEC trace and compared against LUV, GDS, Hyper-G, LRU, and Hybrid algorithms.

D. Least Unified Value (LUV) Algorithm

The LUV algorithm is a more complex functional algorithm that “trie[s] to get the benefit of both LRU and LFU in one unified scheme [3]”. Each object in the cache is assigned a value that is used during a replacement operation; the object with the lowest value is removed. Values for each object (i) are assigned by the following formula:

$$V_i(k) = \frac{C_i}{s_i} \sum_{n=1}^k \left(\frac{1}{2}\right)^{\lambda \Delta_{k,i}}$$

Where: C_i is the cost of object i ; s_i is the size of object i ; λ is a static parameter in the range $0 \leq \lambda \leq 1$; $\Delta_{k,i}$ is the time since the k^{th} reference to object i .

Bahn et al. did not describe a mechanism for determining λ , simply mentioning training as an approach without providing implementation details [5], while Katsaros and Manolopoulos suggested trial-and-error [9]. An obvious

drawback to either methodology is that the resulting parameter may not provide efficient object replacement as request patterns change. Interestingly, a value of zero for λ causes the algorithm to behave very similar to LFU, while behavior similar to LRU results when λ is set to one.

E. Hybrid Algorithm

Wooster et al., presented a cache policy that utilizes several objects and request traffic statistics in a functional computation that derives a cost (or value) for each member or potential member of the cache [10]. The Hybrid formula is defined as:

$$V_i = (r_{tt_s} + W_b / b_s) \cdot (f_i)^{W_n / s_i}$$

Where: s_i = size of object i ; f_i = total number of object requests for object i ; r_{tt_s} = round trip time from cache to server s ; b_s = bandwidth from cache to server s ; W_b and W_n = tuned parameters.

Balamash and Krunz noted that W_b is tuned based on the “importance of the connection time relative to the connection bandwidth [3]”, while W_n is tuned based on the “importance of frequency information relative to the size of the object [3].” These parameters are static; though they can be tuned for each implementation, they do not change over time within the context of the implementation. Similar to LUV, the authors do not provide a methodology for determining the static parameters W_b and W_n other than experimentation utilizing trace files, which can result in less efficient cache performance.

F. Mix Algorithm

Niclausse et al. presented an extension to the Hybrid cache policy that adds a parameter of time since last access to the functional cost computation [7]. The Mix formula is defined as:

$$V_i = \frac{d_i^{r_1} \cdot f_i^{r_2}}{tref_i^{r_3} \cdot s_i^{r_4}}$$

Where: s_i = size of object i ; f_i = total number of object requests for object i ; d_i = average server retrieval delay for object i ; $tref_i$ = current date and time of last request for object i ; r_1, r_2, r_3, r_4 = tuned parameters.

The authors noted that tuning the parameters utilized by the algorithm is not a trivial task and did not present a methodology for doing so. However, their trace simulation experiments found that using a value for r_1 that was much smaller than $r_2, r_3,$ and r_4 gave optimal performance, and their published results used 1, 0.1, 0.1, and 0.1 for the respective parameters. Though the algorithm utilizes several performance characteristics as well as tuned parameters in an attempt to create an efficient replacement algorithm, experimentation by Balamash found that the algorithm was one of the worst performers, generally bested by even LRU and being superior only to Size.

G. Greedy Dual Size with Frequency (GDSF) Algorithm

Jin and Bestavros proposed the GDSF algorithm, which is a functional algorithm similar in approach to Mix in that it utilizes several object statistics, but different in that it utilizes

only one pre-tuned parameter [11]. The GDSF formula is defined as:

$$V_i = f_i \cdot c_i / s_i + L$$

Where: s_i = size of object i ; f_i = total number of requests for object i ; c_i = the cost of bringing object i into the cache; L = a runtime factor which starts at zero when the cache is initialized and represents the value of the most recent object to be replaced.

Arlitt et al. noted that the best HR is achieved when c_i is set to one [12]. The runtime factor L works as follows: when the cache is first initialized, L is set to zero. L remains zero until the cache becomes full and an object needs to be removed from the cache. The algorithm determines the object with the lowest value, V_i , and removes it from the cache. The runtime parameter L is set to the value of V_i for the ejected object. This process continues throughout the life of the cache such that L is an ever-increasing parameter.

Shi and Zhang, attempting to find single optimal algorithms for HR, BHR, and LR separately, instead found that GDSF performed best for all three metrics simultaneously when compared to LRU, LFU, and GDS [13].

H. Lowest Relative Value (LRV) Algorithm

After extensive analysis of web traces, Rizzo and Vicisano proposed a complex (and, according to Bahn et al., “difficult” to implement) algorithm called Lowest Relative Value [14][3]. The algorithm utilizes separate computations based on whether the object is being requested for the first time or a subsequent time:

$$V_i = \begin{cases} P_1(s_i) \cdot [1 - D(t)] \cdot c_i / s_i, n = 1 \\ P_n \cdot [1 - D(t)] \cdot c_i / s_i, otherwise \end{cases}$$

Where: s_i = size of object i ; c_i = the cost of loading object i into the cache; t = time since last request for object i ; P_n = probability of access of $n+1$ given access of n times; $D(t)$ = cumulative distribution function of object inter-access time.

The probability functions P_n and $D(t)$ are based on “extensive analysis of trace data [5],” but are computed dynamically [14]. Rizzo provides an estimation for computing $D(t)$ as: $D(t) \approx c \cdot \log(\frac{t+\tau_1}{\tau_1})$ where τ_1 is a parameter that “accounts for the periodicity of frequent references to popular documents [14]” and c is further defined by the equation: $c = D(t_b) \cdot \log(\frac{t_2+\tau_1}{\tau_1})$

While Bahn, et al. did not examine the effectiveness of LRV, their treatment of the algorithm noted that the V_i equation needed to be calculated for every object in the cache each time an object was removed. Additionally, the LRV authors stated that the probability functions were iterative in nature. While the algorithm is adaptive to traffic, then, it is not only difficult to implement but also computationally complex.

I. Advanced Replacement Cache (ARC) Algorithm

Megiddo and Modha constructed a dynamically adaptive cache policy that attempted to strike a balance between object request frequency and recency. Their policy implements two

LRU cache areas: one is a list of LRU-ordered objects that have been requested only once and the other is a list of LRU-ordered objects that have been requested two or more times [15]. They define their algorithm as

ARC(c) Initialize $T_1 = B_1 = T_2 = B_2 = 0, p = 0$.
 x - requested page.

Case I. $x \in T_1 \cup T_2$ (a hit in ARC(c) and DBL(2c)):
 Move x to the top of T_2 .

Case II. $x \in B_1$ (a miss in ARC(c), a hit in DBL(2c)):
 Adapt $p = \min\{c, p + \max\{|B_2|/|B_1|, 1\}\}$.
 REPLACE(p).
 Move x to the top of T_2 and place it in the cache.

Case III. $x \in B_2$ (a miss in ARC(c), a hit in DBL(2c)):
 Adapt $p = \max\{0, p - \max\{|B_1|/|B_2|, 1\}\}$.
 REPLACE(p).
 Move x to the top of T_2 and place it in the cache.

Case IV. $x \in L_1 \cup L_2$ (a miss in DBL(2c) and ARC(c)):
 case (i) $|L_1| = c$:
 if $|T_1| < c$ then delete the LRU page of B_1 .
 REPLACE(p).
 else delete LRU page of T_1 and remove it from the cache.

case (ii) $|L_1| < c$ and $|L_1| + |L_2| \geq c$:
 if $|L_1| + |L_2| = 2c$ then delete the LRU page of B_2 .
 REPLACE(p).
 Put x at the top of T_1 and place it in the cache.

Subroutine REPLACE(p)

if $(|T_1| \geq 1)$ and $((x \in B_2$ and $|T_1| = p)$ or $(|T_1| > p))$ then
 move the LRU page of T_1 to the top of B_1 and remove it from the cache.

else move the LRU page in T_2 to the top of B_2 and remove it from the cache.

In their implementation, T_1 and T_2 represent the list of most recently requested objects in lists L_1 and L_2 , respectively. Similarly, B_1 and B_2 represent the list of least recently requested objects in lists L_1 and L_2 , respectively. The parameter p dynamically adjusts such that the overall cache contains “the p most recent [objects] from L_1 and $c-p$ most recent [objects] from L_2 .” [15]. The authors’ experimental results show that the ARC algorithm does outperform LRU in their trials. Unfortunately, the experiments were performed using traffic traces specific to their research facility rather than publicly available traces, making direct comparisons to other published algorithms impossible.

III. ADAPTIVE REPLACEMENT POLICY

Memory-based caching being finite in nature is typically governed by a single cache-replacement policy. A simplified and generalized architecture for a cache area x governed by replacement policy R is depicted in figure 1.

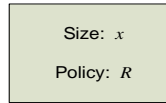


Fig. 1. Simple Cache Architecture

This article focuses on an adaptive cache replacement policy that allows for a short period of possibly less efficient cache performance while still providing some cache benefits, directly leading to an efficient algorithm choice without the need for trace file collection or offline training. The approach is similar to that of ARC, except that it is adaptable to replacement policy choices other than LRU and possibly extensible to more than two policies. The rest of this section outlines the new multiple-algorithm approach to dynamically choosing an efficient replacement policy.

The new approach modifies the simple cache architecture by splitting its finite area into n separate parts, each governed by a distinct replacement policy, R_i . In such an architecture, each partition would start with size x/n . Figure 2 depicts this generalized architecture.

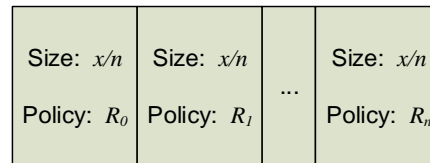


Fig. 2. Generalized n -Policy Cache Architecture

Considering a two-parts cache architecture, a second modification allows for the size of each area y and y' to adjust dynamically based on traffic patterns while their combined sizes are never greater than x , as illustrated in figure 3. Caches are initialized to be completely empty, so a cache system has no history on which to base the loading or removal of objects. The multi-algorithm approach *primes* each cache area with objects as they are being requested. The priming process alternates loading objects between the two caches, y and y' , until both areas are full. Additionally, a cached object can exist in only one area at a time, not both simultaneously. During the load process, the cache policies continue to work normally – if a cached object is re-requested, it is provided by the cache area from which it resides. If one area becomes full during the priming phase, the other area continues to be filled until it, too, is completely primed.

Once primed, a secondary algorithm begins to dynamically adjust the size of the caches based on which area is serving the most cache hits. This is accomplished using a dynamic parameter ρ that is initially set to $x/2$ and is used to calculate the current size of each cache area, mathematically:

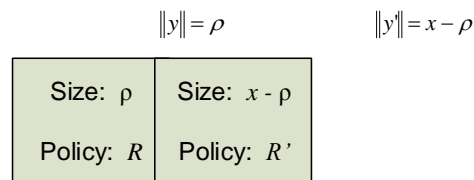


Fig. 3. Two-Policy Cache Architecture with Parameter

When area y experiences a cache hit, ρ is increased by the size of the object being requested. Alternatively, when area y' sees a hit, ρ is decreased by the requested object size. If the total size of all objects cached for an area exceeds the dynamic size, cache replacement policies kick in and objects are removed until the total object size is less than or equal to the dynamic allocation.

The initial theory behind the new multi-algorithm replacement policy hypothesized that a cache area x would adapt to changing traffic patterns in a manner similar to that of ARC. Instead, another interesting result emerged from the simulation results: the overall area x generally converges to either replacement policy R or R' , depending on which policy yields best results for the traffic pattern. The multi-algorithm policy is outlined as follows:

Given:

L_1 is list using LRU policy
 L_2 is list using LFU policy
 x is total cache size
 $0 < |L_1| \leq \rho$ and $\rho < |L_2| \leq x$

LFRU3 (x) Initialize $L_1 = L_2 = 0$, $\rho = x/2$.
 o - requested object.

Case I. $o \in L_1$ (a miss in LFU, a hit in LRU):
If L_1 is primed, L_2 is primed, $\rho > \text{sizeof}(o)$, and $\rho + \text{sizeof}(o) < x$
Adapt $\rho = \rho + \text{sizeof}(o)$.
Move o to the top of L_1 .
If $|L_2| > x - \rho$, evict LFU objects until $|L_2| \leq x - \rho$ and mark L_2 as primed.

Case II. $o \in L_2$ (a miss in LRU, a hit in LFU):
If L_1 is primed, L_2 is primed, $x - \rho > \text{sizeof}(o)$, and $\rho - \text{sizeof}(o) > 0$
Adapt $\rho = \rho - \text{sizeof}(o)$.
Increase frequency of o in L_2 .
If $|L_1| > \rho$, evict LRU objects until $|L_1| \leq \rho$ and mark L_1 as primed.

Case III. $o \notin L_1 \cup L_2$ (a miss in LRU and LFU):
If ((L_1 is not primed and L_1 has fewer entries than L_2)
or (L_1 is not primed and L_2 is primed)
or (L_1 is primed and L_2 is primed and $\rho < x/2$)
and $\rho + \text{sizeof}(o) < x$
(LRU is doing better or needs primed)
add o to the top of L_1 .
Else
If ((L_2 is not primed and L_2 has fewer entries than L_1)
or (L_2 is not primed and L_1 is primed)
or (L_1 is primed and L_2 is primed))
and $\rho - \text{sizeof}(o) > 0$
(LFU is doing better or needs primed)
add o to L_2 with frequency of 1.
 ρ remains unchanged.
If $|L_1| > \rho$, evict LRU objects until $|L_1| \leq \rho$ and mark L_1 as primed.
If $|L_2| > x - \rho$, evict LRU objects until $|L_2| \leq x - \rho$ and mark L_2 as primed.

IV. PRELIMINARY RESULTS

An experiment was designed to assess the effectiveness of the adaptive policy using two algorithms where the cache was split into two parts. LRU algorithm was chosen for policy R and the LFU algorithm was chosen for policy R' . These

policies were chosen due to their computational simplicity [16] and the fact that many if not most existing policies utilize recency and/or frequency in their design [3].

Two traces from the Internet Traffic Archive [17] were chosen to carry out the validation process: one from Digital Equipment Corporation (DEC) and another from the Environmental Protection Agency (EPA). The trace files were chosen primarily based on how they performed under simulation; DEC generally performed better using the LRU policy whereas EPA generally showed a preference for the LFU policy. Additionally, the DEC trace files were previously utilized in algorithm research by in [5], [16], [11], [3], and [14] while the EPA trace files were used in [18]. Finally, a third trace file was created to alternate requests from the DEC and EPA trace files in order to craft a synthetic trace file for analysis that mimicked two simultaneous unique traffic patterns.

These web traces were applied using a cache simulator written by Pei Cao, known as Uniform [19]. Employed in other research ([20], [21], and [22]) and written in C, this application readily facilitated the simulation of the LRU replacement policy. The simulator was enhanced by implementing the LFU policy as well as the adaptive policy, LFRU3, in order to simulate these policies along with the already available LRU. Additionally, each policy utilized a threshold mechanism that existed in the initial simulator and was implemented in newly added policies (LFU and LFRU3). The threshold process refused to cache objects above a certain size (250kB in this experiment) so that one large object could not 'pollute' the cache, meaning that a request for one large object would not cause many smaller and possibly more beneficial objects to be removed.

The performance of the three primed with thresholding versions of the three algorithms were compared to determine the effectiveness of LFRU3. Figures 4-6 depict the performance using the DEC, EPA, and synthetic DECEPA trace files, respectively. For the DEC trace the adaptive policy, LFRU3, has nearly the same performance as the best algorithm (LRU) for smaller cache sizes and very close in performance at the largest cache size, as shown in figure 4. This indicates that LFRU3 successfully converged to the better of the two algorithms in all test cases.

In the EPA simulations the new two-algorithm policy chose the better (LFU) algorithm in two scenarios – cache sizes of 0.05% and 10% of maximum (5 simulations). For these five simulations, then, LFRU3 successfully converged to the better of the two algorithms 40% of the time. It is encouraging to note that at the 0.05% cache size, where the LRU algorithm and LFU algorithm showed the most dramatic performance differential, the adaptive policy successfully converged to the better of the two algorithms, as shown in figure 5.

The synthetic simulation showed the most promise. In this simulation the new approach was actually superior to the other algorithms for smaller cache sizes. It delivered an 8% HR improvement over LFU and 48% improvement over LRU for the 0.05% cache size. At 0.5% cache size, the results were less pronounced but still superior: 1.4% HR improvement

over LFU and 9.6% better than LRU, as shown in figure 6. Figures 7 and 8 depict the HR for these two cache sizes over simulation time.

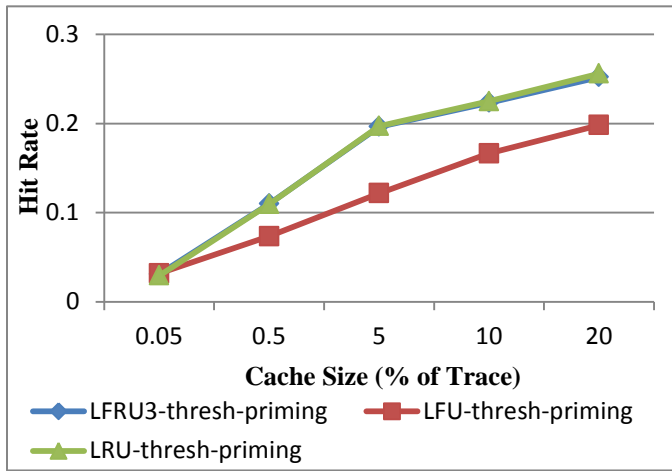


Fig. 4. HR vs. Cache Size (percentage of DEC Trace)

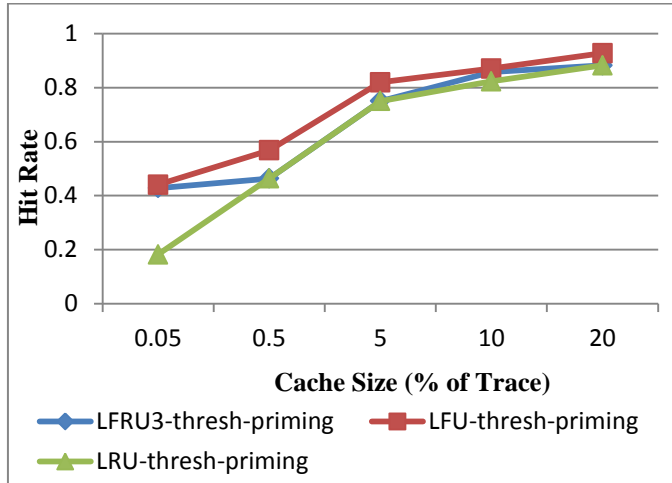


Fig. 5. HR vs. Cache Size (percentage of EPA Trace)

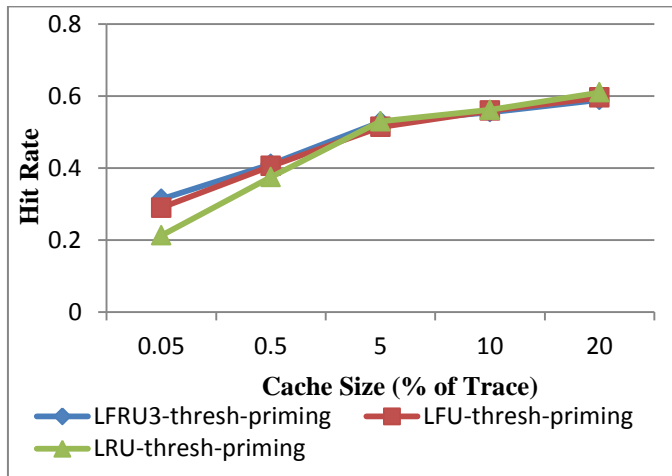


Fig. 6. HR vs. Cache Size (percentage of DECEPA)

Results in Figure 7 illustrate that for the smallest cache, LFRU3 performs consistently better than the other algorithms, and all three algorithms have consistent performance throughout the life of the simulation. Thus, the convergence to superior performance over the other algorithms occurs very early in the lifecycle of the cache and remains constant throughout, a very encouraging result.

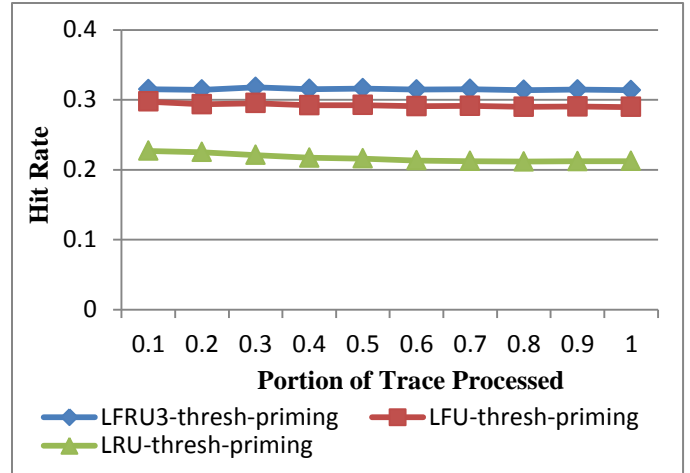


Fig. 7. HR vs. Sim. Time Size (0.05% Synthetic Trace)

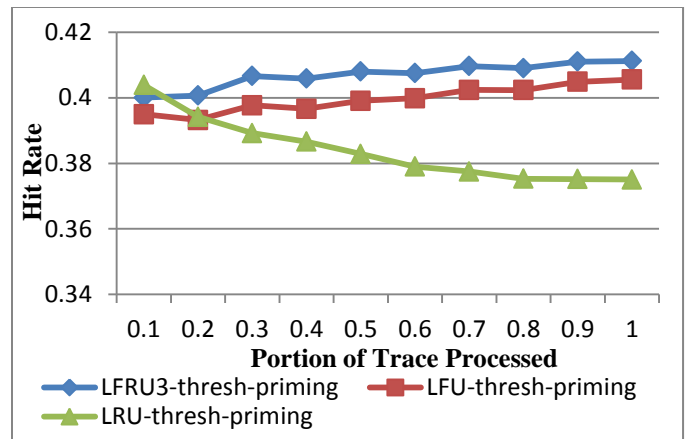


Fig. 8. HR vs. Sim. Time Size (0.5% Synthetic Trace)

The next largest cache shows LRU starting strong but quickly being surpassed by both LFU and the new approach, with LFRU3 consistently better than LFU throughout the examination period, as shown in figure 8. While in this case the new policy is not superior at the first examination period (10% of the trace), it quickly exceeds the performance of the initial best performer and shows gradual improvement and superior performance throughout the remainder of the cache lifecycle.

For Byte Hit Rate (BHR), the performance gain was not as pronounced. As Figure 9 illustrates, the LFRU3 policy slightly edges the next-best performer, LFU, at the smaller size, but performance at other sizes varies. Reduced latency (LR), a measure of the reduction of time spent waiting for objects, shows promise.

Figure 10 shows that LR is nearly identical to that for HR – LRFU3 is the best LR performer for the two smallest caches.

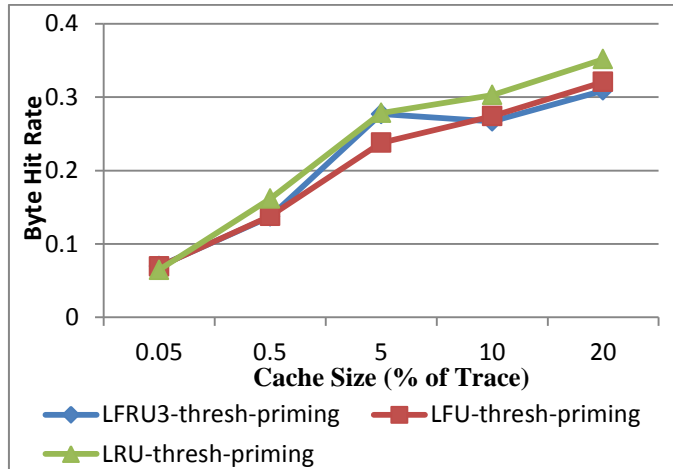


Fig. 9. BHR vs. Cache Size (percentage of Striped Trace)

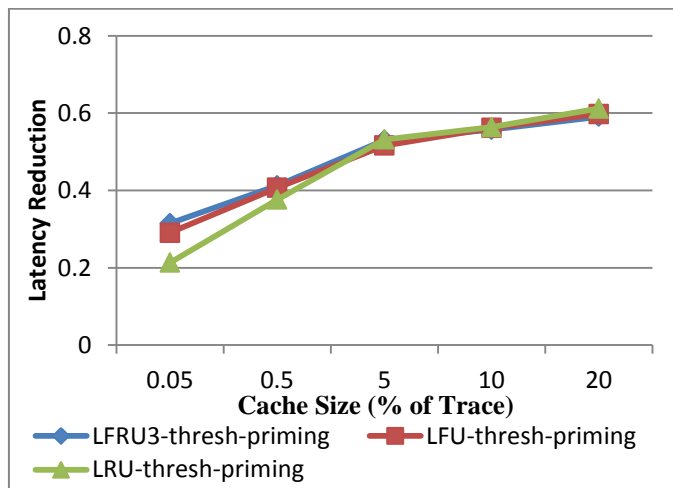


Fig. 10. LR vs. Cache Size (percentage of Striped Trace)

V. CONCLUSIONS

Caching for Web documents is a hugely beneficial function, and much research has been published on cache replacement policies. While the replacement algorithms themselves may factor in one, a few, or many parameters related to the objects and their request history, a multi-algorithm policy has not been attempted. While this article focuses on a two-policy architecture, it is expected that it can be effective for more than two policies.

Simulations of the multi-algorithm cache replacement policy shows that it is a viable approach that can adapt itself to the better of two replacement policies in many instances, and provide superior performance in some others. The empirical results support that the adaptive policy works particularly well in environments with limited cache sizes.

REFERENCES

- [1] Smith, A. J., "Bibliography on Paging and Related Topics", Operating Systems Reviews, Vol. 12, 1978.
- [2] Smith, A. J., "Second Bibliography for Cache Memories", Computer Architecture News, Vol. 19, No. 4, 1999.
- [3] Balamash, A. and M. Krunz, "An Overview of Web Caching Replacement Algorithms", IEEE Communications Surveys, Vol. 6, No. 2, Second Quarter, 2004.
- [4] Baeza-Yates, Ricardo, et al., "Design Trade-Offs for Search Engine Caching", ACM Transactions Web, Vol. 2, No. 4, October, 2008.
- [5] Bahn, Hyokyung, et al., "Efficient Replacement of Nonuniform Objects in Web Caches", IEEE Computer, Vol. 35, No. 6, June, 2002.
- [6] Chankhunthod, A., et al., "A Hierarchical Internet Object Cache", Technical Report 95-611, Computer Science Department, University of Southern California, Los Angeles, California, 1995.
- [7] Niclausse, N., et al., "A New and Efficient Caching Policy for the World Wide Web", Proceedings of Workshop Internet Server Performance (WISP 98), 1998.
- [8] Williams, S., "Removal Policies in Network Caches for World Wide Web Documents", Proceedings of ACM SIGCOMM Conference, Stanford University, Aug. 1996.
- [9] Katsaros, D. and Y. Manolopoulos, "Caching in Web Memory Hierarchies", Proceedings of the 2004 ACM Symposium on Applied Computing, 2004.
- [10] Wooster, R. and M. Abrams, "Proxy Caching that Estimates Page Load Delays", Proceedings of the 6th International World Wide Web Conference, Santa Clara, CA, Apr. 1997.
- [11] Shudong and A. Bestavros, "Popularity-Aware GreedyDual-Size Web Proxy Caching Algorithms", Proceedings of the 20th IEEE International Conference on Distributed Computing Systems (ICDCS 2000), 2000.
- [12] Arlitt, M., et al., "Evaluating Content Management Techniques for Web Proxy Caches", SIGMETRICS Performance Evaluation Review, Vol. 27, No. 4, March, 2000.
- [13] Shi, Lei and Y. Zhang, "Optimal Model of Web Caching", Conference Record of the 2008 Fourth International Conference on Natural Computation, 2008.
- [14] Rizzo, L. and L. Vicisano, "Replacement Policies for a Proxy Cache", IEEE/ACM Transactions on Networking, Vol. 8, No. 2, 2000.
- [15] Megiddo, N. and D. Modha, "Outperforming LRU with an Adaptive Replacement Cache Algorithm", IEEE Computer, Vol. 37, No. 4, April, 2004.
- [16] Bahn, Hyokyung, "Web Cache Management Based on the Expected Cost of Web Objects", Information and Software Technology, Vol. 47, April, 2005.
- [17] The Internet Traffic Archive, <http://ita.ee.lbl.gov/html/traces.html>, last updated April 9, 2008.
- [18] Cheng, A. M. K. and Z. Zhang, "Improving Web Server Performance with Adaptive Proxy Caching in Soft Real-time Mobile Applications", Journal of VLSI Signal Processing, Vol. 47, 2007.
- [19] Cao, Pei, Uniform Cache Simulator, download from <ftp://ftp.cs.wisc.edu/pub/cao/webcache-simulator.tar.z>.
- [20] Cao, P. and S. Irani, "Cost Aware WWW Proxy Caching Algorithms", Proceedings of the 1997 UXENIX Symposium on Internet Technology and Systems, 1997.
- [21] Shi, Lei, et al., "An Applicative Study of ZipF's Law on Web Cache", International Journal of Information Technology, Vol. 12, No. 4, 2006.
- [22] Buijtenjijk, V., "Module Deployment and Management within the Grid-based Virtual Laboratory for e-Science", Thesis Paper, Universiteit van Amsterdam, May, 2005.

Application of Fuzzy Self-Optimizing Control Based on Differential Evolution Algorithm for the Ratio of Wind to Coal Adjustment of Boiler in the Thermal Power Plant

Ting Hou

College of Electronic and Electrical
Engineering
Shanghai University of Engineering
Science
Songjiang District, Shanghai
201620, China

Liping Zhang*

College of Electronic and Electrical
Engineering
Shanghai University of Engineering
Science
Songjiang District, Shanghai
201620, China

Yuchen Chen

College of Electronic and Electrical
Engineering
Shanghai University of Engineering
Science
Songjiang District, Shanghai
201620, China

Abstract—The types of coal are multiplex in domestic small and medium sized boilers, and with unstable ingredients, the method of maintaining the amount of wind and coal supply in a fixed proportion of the wind adjustment does not always ensure the best economical boiler combustion process, the key of optimizing combustion is to modify reasonable proportion of wind and coal online. In this paper, a kind of fuzzy self-optimizing control based on differential evolution algorithm is proposed, which applied in the power plant boiler system, the boiler combustion efficiency has been significantly improved than previous indirect control. In this paper, a thermal power plant is our research object, in the case of determining the optimum system performance, the unit efficiency can be increased significantly using this method, and the important issues of energy efficiency of power plants can be successfully solved.

Keywords—fuzzy self-optimizing control; differential evolution algorithm; best ratio of wind to coal; boiler efficiency

I. INTRODUCTION

Boiler combustion control system is a nonlinear system, and with the characteristics of time-varying, noise disturbing, pure hysteresis, the experience of the operator can't be used with the conventional control methods. But in the process of the actual operation of the boiler, these experiences information is important. Optimal control of modern control theory is also difficult to be applied in the modelling of such systems. However, fuzzy control can be just applied to control the unknown or changing process for this kind of mathematical model [1]. Using fuzzy theory and fuzzy control method, the experience of the operator can be summed up. Moreover, a kind of fuzzy control strategy can be achieved using this fuzzy logic inference method.

In this paper, thermal power plant is our research object, the power plant is small, and it has three 12MW units, 3 sets of 150t/h coal-fired boiler. The self-optimizing simulation is used to find the best ratio of wind to coal, and combining differential evolution algorithm, the quantization factor K_y will be

optimized, then fuzzy control technology is used to regulate the speed of blower speed and coal feeder, finally 3 boilers of the thermal power plant cogeneration combustion efficiency significantly will be improved than previous indirect control, which is agreed with our above analysis.

II. BOILER FUZZY SELF-OPTIMIZING CONTROL SYSTEM AND DIFFERENTIAL EVOLUTION ALGORITHM

A. PID fuzzy controller system

The error and error change rate between the amount transferred and the fixed value as input variables are always used in the general fuzzy controller; therefore, it has the similar effect with the conventional PD controller^[2]. A good dynamic quality can be obtained using such fuzzy controller in control system. However, steady-state error of the amount transferred is difficult to eliminate. In order to eliminate the steady-state error of the control system, PID fuzzy control system can be used as shown in Fig1. In the figure, e and \dot{e} are respectively the error and error derivative. Conventional control table form is used for fuzzy control.

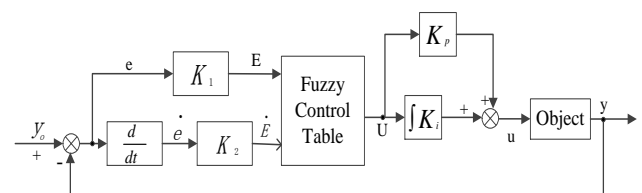


Fig. 1. PID fuzzy controller schematics

Output control table need to be changed to the amount of actual control value through an output link, and then coupled to the controlled object to achieve the control. Commonly, the two ways of output link used are proportional output and integral output, the former has a fast step response characteristic, but it is error control; the latter can be close to the no error control, but with the slow response and the

overshoot is relatively large. The combination of both ways is used in the system, which have advantages of a small overshoot and short time transient.

Mathematical expression of PD-type fuzzy controller can be derived by algebraic product-addition-focus fuzzy inference method:

$$u = A + Pe + D\dot{e} \quad (1)$$

$$\text{In formula (1), } A = u_{i,j} - \frac{u_{i+1,j} - u_{i,j}}{e_{i+1} - e_i} e_i - \frac{u_{i+1,j} - u_{i,j}}{\dot{e}_{j+1} - \dot{e}_j} \dot{e}_j$$

$$D = \frac{u_{i,j+1} - u_{i,j}}{\dot{e}_{j+1} - \dot{e}_j}$$

$$P = \frac{u_{i+1,j} - u_{i,j}}{e_{i+1} - e_i}$$

Thus input-output relationship of the input and output PID fuzzy controller can obtained as shown in Figure 1:

$$\begin{aligned} u &= K_p U + K_i \int U dt = \\ &= K_p (A + PK_1 e + DK_2 \dot{e}) + \\ &\quad K_i \int (A + PK_1 e + DK_2 \dot{e}) dt \\ &= K_p A + K_i At + (K_p K_1 P + K_i K_2 D) \\ &\quad + K_i K_1 P \int e dt + K_p K_2 D \dot{e} \end{aligned} \quad (2)$$

The above mentioned fuzzy controller have four adjustable parameters: quantify factors K_1 、 K_2 , scale factor K_p and integral coefficient K_i , increasing K_1 、 K_2 , the resolution of the error can be improved, then the control accuracy is also improved. But if K_1 、 K_2 are too big, it will bring unstable for the system^[3]. Increasing K_p or K_i , response speed will increase, but which will lead to oscillate. Based on actual adjusted experience, the desirable value as $K_1 \approx K_2$, $K_p = (2 \sim 3) K_i$. When K_1 and K_2 are too larger, K_p and K_i should be decreased, and when the sampling period is longer, then K_p and K_i can be chosen larger.

B. Fuzzy self-optimizing controller in the ratio of wind to coal

In the boiler control, due to the amount of pulverized coal cannot be accurately measured online, so the excess air coefficient α is used to substitute β which represents the ratio of wind to coal volume, so ratio of wind to coal and α are equivalent. The ratio of actual air volume (V) and the theoretical air amount (V_0) is the excess air factor α . According to "Safety Engineering Dictionary", exhaust of the boiler is qualified when the excess air coefficient α is within the range of 1.15 to 1.25 of the thermal power plant^[4], in addition, the thermal efficiency is highest at this time. If excess air coefficient α is too large, excess air flue gas will take too much heat, increasing exhaust gas temperature will lead to heat loss q_2 increasing, at the same time, it will produce a large amount of NO_x and SO_x pollution; on the contrary, if α is too small, complete combustion of fuel cannot be guaranteed. Selecting

the appropriate α is an important means to reduce boiler heat loss and improve thermal efficiency.

In this system, α is the seeking optimization index, if:

$$\alpha \notin [1.15, 1.25] \quad (3)$$

Start fuzzy self-optimizing search, when α is within the range of 1.15 to 1.25, indicates that the optimum working range has been found, and then optimization can be stopped.

Self-optimizing fuzzy controller works as follows: combustion coal as an index used to find the best ratio of wind to coal^[5] with fuzzy self-optimizing controller. Increment of coal consumption Δy will be measured at each sampling period, accord to Δy and the optimization step of previous cycle to determinate the optimization step. ΔY and ΔX are coal consumption and fuzzy linguistic variables of the steps. K_y is quantization factor of Δy , which converts the fuzzy linguistic variables of coal consumption increment ΔY to the actual value, K_x is scale factor, which convert the ΔX to actual value of the step.

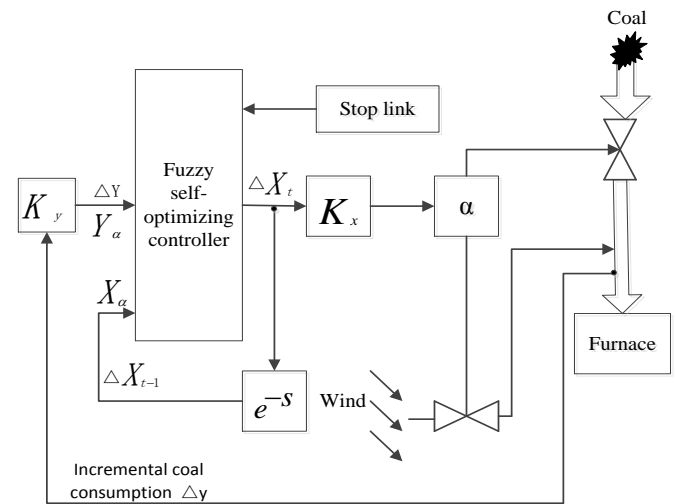


Fig. 2. Fuzzy self-optimizing controller based on differential evolution algorithm for the ratio of wind to coal

In practical applications, in order to ensure the stability of self-optimizing process, a stop link^[6] is added. If the temperature of the furnace has a big fluctuation due to environment distribution, the search should be stopped to avoid malfunction.

Selecting ΔY , ΔX respectively as 8 and 6 linguistic variables of fuzzy subsets is contained as follows:

$$\Delta Y = \{NB, NM, NS, NO, PO, PS, PM, PB\}$$

$$\Delta X = \{NB, NM, NS, PS, PM, PB\}$$

NB, NM, NS, NO, PO, PS, PM, PB respectively denote negative big, negative medium, negative small, negative zero, positive zero, positive small, positive medium and positive big.

The fuzzy domain of ΔY and ΔX are defined as 14 and 12 grades:

$$Y_{\alpha} = \{-6, -5, -4, -3, -2, -1, -0, +0, +1, +2, +3, +4, +5, +6\}$$

$$X_{\alpha} = \{-6, -5, -4, -3, -2, -1, +1, +2, +3, +4, +5, +6\}$$

Self-optimizing search process control rules showed as in Table I. ΔX_{t-1} is the optimization step of previous cycle, ΔX_t is the optimization step. Table II shows the rule table about fuzzy self-optimizing control in the ratio of wind to coal.

TABLE I. RULE TABLE ABOUT FUZZY SELF-OPTIMIZING CONTROL IN THE RATIO OF WIND TO COAL

| ΔX_t \ ΔX_{t-1} \ ΔY | NB | NM | NS | PS | PM | PB |
|--|----|----|----|----|----|----|
| NB | PB | PB | PB | NB | NB | NB |
| NM | PM | PB | PB | NB | NB | NM |
| NS | PS | PM | PM | NM | NM | NS |
| NO | PS | PS | PS | NS | NS | NS |
| PO | PS | PS | PS | NS | NS | NS |
| PS | NS | NM | NM | PM | PM | PS |
| PM | NM | NB | NB | PB | PB | PM |
| PB | NB | NB | NB | PB | PB | PB |

Control strategy of the controller can be summarized as follows:

IF $\Delta X_{t-1} = \text{NB}$ AND $\Delta Y = \text{NB}$ THEN $\Delta X_t = \text{PB}$

IF $\Delta X_{t-1} = \text{NB}$ AND $\Delta Y = \text{NM}$ THEN $\Delta X_t = \text{PM}$

...

Apply synthesis rules of fuzzy inference, and coupled with manual correction, the ultimate self-optimizing control table can be obtained as shown in Table II.

TABLE II. TABLE ABOUT FUZZY SELF-OPTIMIZING CONTROL IN THE RATIO OF WIND TO COAL

| X_{at} \ X_{at-1} \ Y_{α} | -6 | -5 | -4 | -3 | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 |
|--------------------------------------|----|----|----|----|----|----|----|----|----|----|----|----|
| -6 | 6 | 6 | 6 | 6 | 6 | 6 | -6 | -6 | -6 | -6 | -6 | -6 |
| -5 | 6 | 6 | 6 | 6 | 6 | 6 | -6 | -6 | -6 | -6 | -6 | -6 |
| -4 | 4 | 4 | 5 | 5 | 6 | 6 | -6 | -6 | -5 | -5 | -4 | -4 |
| -3 | 4 | 4 | 5 | 5 | 6 | 6 | -6 | -6 | -5 | -5 | -4 | -4 |
| -2 | 1 | 1 | 3 | 3 | 4 | 4 | -4 | -4 | -3 | -3 | -1 | -1 |
| -1 | 1 | 1 | 3 | 3 | 4 | 4 | -4 | -4 | -3 | -3 | -1 | -1 |
| -0 | -1 | -1 | -1 | -1 | -1 | -1 | 1 | 1 | 1 | 1 | 1 | 1 |
| +0 | 1 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | -1 | -1 |
| 1 | -1 | -1 | -3 | -3 | -4 | -4 | 4 | 4 | 3 | 3 | 1 | 1 |
| 2 | -1 | -1 | -3 | -3 | -4 | -4 | 4 | 4 | 3 | 3 | 1 | 1 |
| 3 | -4 | -4 | -5 | -5 | -6 | -6 | 6 | 6 | 5 | 5 | 4 | 4 |
| 4 | -4 | -4 | -5 | -5 | -6 | -6 | 6 | 6 | 5 | 5 | 4 | 4 |
| 5 | -6 | -6 | -6 | -6 | -6 | -6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | -6 | -6 | -6 | -6 | -6 | -6 | 6 | 6 | 6 | 6 | 6 | 6 |

The search speed can be improved by increasing K_x and K_y , the value of K_x can also affect the loss of search, so K_y can be chosen based on the search speed requirements, K_x can be chosen according to the requirements of the loss of search. The differential evolution algorithm is added in the system to optimize the value of the quantization factor K_y , making the fuzzy value of incremental coal consumption can be conversed

more precise to the actual value after each optimization, boiler efficiency ratio is improved significantly, and especially at low load.

C. Differential evolution algorithm

The basic idea of differential evolution algorithm is: the individuals of current population followed through mutation and crossover operation and produce tested individual, then, based on the greedy thoughts, the optimum individual is selected; thereby the new populations are generated. When in the process of t times iteration, the i -th population of individuals $x_i^t = (x_{i1}^t, x_{i2}^t, \dots, x_{id}^t), i=1, 2, \dots, N_p$ is a d -dimensional candidate solution vector, the individual with best fitness value of all individuals are denoted $x_{best}^t = (x_{best1}^t, x_{best2}^t, \dots, x_{bestd}^t)$, then three-steps operation as mutation, crossover, selection are performed in the process of the t -th iterations^[7].

1) Mutation

Mutation operation is achieved by DE algorithm with the differential method. There are a variety of differential strategies, common differential strategy is to select randomly two different individuals in the t -th generation of the population, after multiplying the vector difference of the two individual with variability factor, and a synthetic new individuals with upcoming individual variation u_j^t, u_j^t is called variation vector. DE/rand/1 variation formula as follows:

$$u_{id}^t = x_{ij}^t + F \cdot (x_{r2j}^t - x_{r3j}^t) (j = 1, 2, \dots, d) \quad (4)$$

There, u_{ij}^t is j -dimensional components of variation vector u_j^t , F is variability factor, used to adjust the differential scaling of multiple, usually selected as the value between (0, 1], three random numbers r_1, r_2, r_3 are different from each other and not equal to i, x_{r1}^t is called a parent basis vectors, $(x_{r2}^t - x_{r3}^t)$ is a parent difference vector, then crossover to x_i^t and variation vector u_j^t are implemented.

2) Crossover

i -th individual x_i^t and variation vector u_j^t generated by mutation work as the following crossover:

$$y_{ij}^t = \begin{cases} u_{ij}^t, & \text{if } \text{rand} \leq C_R \\ x_{ij}^t, & \text{otherwise} \end{cases} (j = 1, 2, \dots, d) \quad (5)$$

There, y_{ij}^t is the individual of x_i^t j -dimensional components corresponding to the test individual y_i^t , rand is random numbers distributed uniformly between [0, 1], C_R is cross-factor, usually the value is pre-set within the range [0, 1], cross factor determines the proportion of each component of variation vector u_j^t in the tested individuals y_i^t . When $C_R=1$, tested individual is equal to the variation individual.

3) Selection

Comparing produced tested individual y_i^t with the individual x_i^t , the optimum individual of both individual is selected as the next generation. Therefore, comparing the fitness of tested individual y_i^t and individual x_i^t , when the fitness of tested individual y_i^t is better than the fitness of individuals x_i^t , y_i^t will replace individual x_i^t and become the next generation of

individuals x_i^{t+1} , otherwise, x_i^t will be retained and become individual x_i^{t+1} of generation $t+1$.

$$x_i^{t+1} = \begin{cases} y_i^t, & \text{if } \text{fitness}(y_i^t) \leq \text{fitness}(x_i^t) \\ x_i^t, & \text{otherwise} \end{cases} \quad (6)$$

When the fitness of x_i^{t+1} is better than the fitness of x_{best}^{t+1} , update x_{best}^{t+1} , make $x_{\text{best}}^{t+1} = x_i^{t+1}$, and save the best fitness value $\text{fitness}_{\text{best}}$. In this way, crossover, mutation and selection operations should be repeated, until the population optimal adaptation meet the pre-set threshold value or reach the maximum number of iterations. At this point the best fitness value corresponding to the individual is the optimal solution searched by DE algorithm for optimizing this problem. The system uses differential evolution algorithm to optimize quantitative factor K_y to improve the efficiency of the boiler system, especially in low-load operation.

III. EFFECT OF THIS SYSTEM AFTER APPLIED TO THE THERMAL POWER PLANT

A. Improvement of boiler efficiency

Boiler efficiency is the average in a certain statistical cycle, the average efficiency of a reference period is calculated as follows:

$$\eta = 9143 \frac{\sum D_N}{\sum B_N} \% \quad (7)$$

In formula (7), $\sum D_N$ represents the standard steam production during the reference period; $\sum B_N$ represents the standard coal consumption during the reference period.

Fuzzy self-optimizing controller in the ratio of wind to coal is used in unit, the coal amount is fed back to input value, change the wind volume to change the ratio of wind to coal to achieve the high efficiency of the boiler, the simulation results shown in Figure 3.

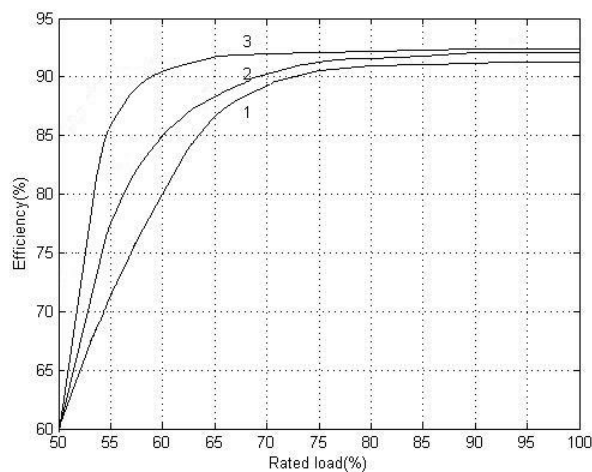


Fig. 3. Contrast curve of boiler efficiency

In figure 3, curve 1 is the boiler efficiency when manually adjusting the ratio of the wind to coal of the system, curve 2 is

boiler efficiency after adding self-optimizing fuzzy control, and curve 3 is boiler efficiency after using differential evolution algorithm to optimize quantitative factors K_y . As seen from figure 3, when running with full capacity, boiler efficiency is improved as approximately 0.9% after using fuzzy self-optimizing controller in the ratio of wind to coal than manual control, and incorporating the differential evolution algorithm into the boiler system, its efficiency can be increased by approximately 0.2%, especially in low-load operation, efficiency is more significant.

B. Saving coal and emission reduction of the power

Restricted by detection technology and indirect control, improper wind supply often causes a substantial loss of heat or incomplete combustion of fuel. The best control system based on the current thermal power plants can only guarantee the energy conversion efficiency is generally about 35% [8], it is already a very high conversion efficiency of burning calories, the average thermal efficiency of thermal power plants is about 32.5%, system efficiency can be increased by 1.1% when running with full capacity, so in this way we can get considerable energy savings.

Note: The standard coal calorific value calculated using 7000 Kcal / kg. The average heating ratio of thermal power plant is about 84.58%.

| | 24h | 30d | 365d |
|-----------------------------|------------------------|--------------------------|---------------------------|
| Output power /kWh | 0.2714×10 ⁶ | 8.143×10 ⁶ | 99.07×10 ⁶ |
| Heat input /Kcal | 158.71×10 ⁷ | 46764.69×10 ⁷ | 60976.937×10 ⁷ |
| Coal consumption of unit /t | 226.72 | 6680.67 | 87109.97 |

| | 24h | 30d | 365d |
|-------------------------------------|------------------------|------------------------|-------------------------|
| Output power /kWh | 0.2743×10 ⁶ | 8.2335×10 ⁶ | 100.16×10 ⁶ |
| Saving energy /Kcal | 4.45×10 ⁷ | 133.04×10 ⁷ | 1624.79×10 ⁷ |
| Saving standard coal consumption /t | 2.494 | 74.789 | 957.167 |

It can be seen from Table III and Table IV, the energy-saving effect is obvious after using the control method in this article.

Combining with fuzzy self-optimizing control method based on differential evolution algorithm in this article, the boiler combustion control technology is achieved, applied in the three 12MW units in the thermal power plant, its efficiency is increased by 1.1%, a boiler can save 957 tons of standard coal one year, so three units can save about 2,871 tons of standard coal in a year. This has a significant effect on energy-saving, environmental protection, and reducing atmospheric dust pollution for the plant.

IV. CONCLUSION

Fuzzy self-optimizing controller in the ratio of wind to coal is used in the unit with appropriate modification, the coal amount is fed back to the forward input value, and so that ventilation can rapidly changing with the amount of fuel changes, and the ratio of wind to coal can be controlled within an appropriate range.

Differential evolution algorithm combined with fuzzy self-optimizing control system enables the boiler operation efficiency significantly improves especially in the situation of low load.

In the case of determining the optimal performance of the system, fuzzy control technology used to make the wind volume can quickly search for the best value, improve the thermal efficiency of the boiler, achieve the effect of economizer and efficiency, and the economics of the thermal power plant is improved greatly.

ACKNOWLEDGMENT

The project has been supported by the Innovation Program of Shanghai Municipal Education Commission "Vehicle

Collision Avoidance System based on Vehicle Wireless Communication" (No.12YZ151).

REFERENCES

- [1] Ji Chang-an et al. Image processing and transformation of power plant application of fuzzy control[J]. High Voltage Engineering, 2007(33-11).
- [2] Mao Pei. Research Minhang Power Plant 125MW unit of energy-saving technology to run[D]. Shanghai Jiaotong University, 2004.
- [3] Bai Ruixiang, Tai Xinmeng, Li Huien. Fuzzy Self-optimizing Control on Combustion System of Industrial Boiler [J]. Instrument Technique and Sensor, 2009, 6 (98-100).
- [4] Cui Keqing. Safety Engineering Dictionary [M]. Beijing: Chemical Industry Press, 1995(342-350).
- [5] Ai Hong. Dynamic Fuzzy External Control of the Efficiency of Boiler Combustion[J]. Techniques of Automation & Applications, 2002, 21(5-8).
- [6] Li Shiyong. Fuzzy control [M]. Harbin: Harbin institute of technology press. 2011(172-177).
- [7] Wu Lianghong, Wang Yaonan, Yuan Xiaofang, Zhang Jian. Research on Differential Evolution Algorithm for MOPs[J]. Journal of Hunan University(Natural Sciences), 2009, 36, 2(53-57).
- [8] Ling Weiping. The use of fuzzy control technology of wind and coal ratio to the design of power plant [J]. Northeast Electric Power Technology, 2011, 7(43-45).

Automatic Optic Disc Boundary Extraction from Color Fundus Images

Thresiamma Devasia
Dept. of Computer Science
Assumption College
Changanacherry, India

Paulose Jacob
Dept. of Computer Science
Cochin University of
Science and Technology
Cochin, India

Tessamma Thomas
Dept. of Electronics
Cochin University of
Science and Technology
Cochin, India

Abstract—Efficient optic disc segmentation is an important task in automated retinal screening. For the same reason optic disc detection is fundamental for medical references and is important for the retinal image analysis application. The most difficult problem of optic disc extraction is to locate the region of interest. Moreover it is a time consuming task. This paper tries to overcome this barrier by presenting an automated method for optic disc boundary extraction using Fuzzy C Means combined with thresholding. The discs determined by the new method agree relatively well with those determined by the experts. The present method has been validated on a data set of 110 colour fundus images from DRION database, and has obtained promising results. The performance of the system is evaluated using the difference in horizontal and vertical diameters of the obtained disc boundary and that of the ground truth obtained from two expert ophthalmologists. For the 25 test images selected from the 110 colour fundus images, the Pearson correlation of the ground truth diameters with the detected diameters by the new method are 0.946 and 0.958 and, 0.94 and 0.974 respectively. From the scatter plot, it is shown that the ground truth and detected diameters have a high positive correlation. This computerized analysis of optic disc is very useful for the diagnosis of retinal diseases.

Keywords—*fundus image; optic nerve head; optic disc; Fuzzy C-Means clustering*

I. INTRODUCTION

The fundus images are used for diagnosis by trained clinicians to check any abnormality or any sort of change in the retina. A healthy retinal image contains anatomical structures like the macula, the optic disc and blood vessels. The retinal optic disc is the region from where the optic nerve of the retina emanates. Hence, it often serves as an important landmark and reference for other features in a retinal fundus image. The optic disc (OD) in a healthy retinal image usually appears as a bright yellowish and circular shaped object which is partly covered with blood vessels. Diseases with symptoms on the fundus images are very complex. For the OD, differences in the color, shape, edge or vasculature may signify a pathological change [1]. The information about the OD can be used to detect the severity of some diseases such as glaucoma.

An efficient segmentation of the OD is essential to diagnose various stages of many retinal diseases. The optic disc appears as an elliptical region with high intensity in retinal images [2]. The method of optic disc boundary detection can be separated into two steps: optic disc extraction and disc boundary detection.

Many existing techniques can be used with reasonable success to extract the optic disc and disc boundary. Huiqi Li et al. [3] designed yet another method based on Principle Component Analysis to localize OD automatically. Snake Active Contour methodology for optical disc detection was proposed by Thitiporn et al. [4]. The contrast of the optical disc is used as the significant feature in this work. But the initialization of size and shape of the contour is in fact, a practical difficulty of this approach.

A novel method to locate the optic nerve in fundus images was developed by Hoover A. et al. [5]. They used Fuzzy convergence to determine the origin of the blood vessel network. Hough Transform was used for OD detection by Chrastek et al. [6]. Juan Xu et al. [7] used a modified Active Contour Model for OD detection. The smoothing update equation of Snake model is modified and used in this approach for performance enhancement. Genetic Algorithm (GA) based optic disk detection is reported by Enrique et al.[8]. Echegaray et al.[9] developed a method for automatic initialization of a level-set segmentation algorithm to find the margin of the optic disc in fundus images as an indicator for glaucoma.

A method to automatically segment the optic using morphological approach was developed by Welfer D et al.[10]. Siddalingaswamy et al. [11] proposed a method for automatic localization and accurate boundary detection of OD using iterative thresholding technique. Yuji Hatanaka, et al. [12] used a Canny edge detection filter for OD edge detection. Amin Deghani et al. [13] designed histogram matching technique for OD localization. Angel Suero et al.[14] used morphological techniques for OD localization. Fuzzy C Means Clustering with thresholding is used in this work for the extraction of optic disc. Fuzzy C-Means is an unsupervised technique that has been successfully applied for feature analysis, clustering, and to classify designs in the fields such as geology, medical imaging and image segmentation.

The rest of the paper is organized as follows. Section II describes the materials used for the new method. In Section III, a new algorithm for the efficient extraction of optic disc boundary ocular fundus images is presented. The results are presented in Section IV, and conclusions are given in Section V.

II. MATERIALS AND METHODS

All the images used in this paper are obtained from the DRION database. There are 110 retinal colour fundus images

with an array size of $400 \times 600 \times 3$ pixels, along with the optic nerve contours traced by two experts.

III. ALGORITHM DEVELOPED

The algorithm is composed mainly of four steps. First, the red channel of the colour retinal image is separated. Then the blood vessels are removed. Thirdly the Fuzzy C Means Clustering operation is applied on the vessel-free image. Thresholding is used for the extraction of OD from this image. The boundary of the extracted OD is traced and, finally this extracted boundary is overlaid on the ground truth image. Performance analysis is done comparing the detected diameters with the ground truth diameters. Statistical analysis is done using scatter plot, which gives high positive correlation between the detected boundary and ground truth boundary.

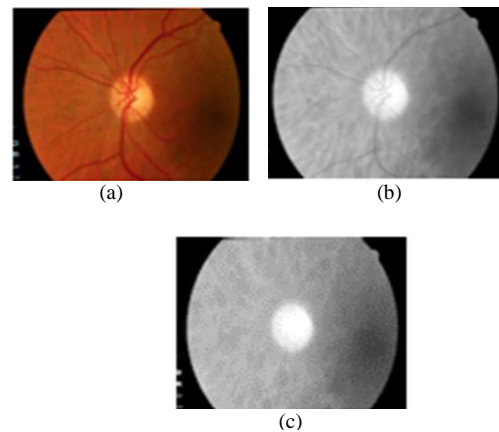


Fig. 1. The preprocessing steps (a) Input Image (b) Red channel (c) Blood vessel Removal

A. Preprocessing

The preprocessing step removes variations due to image acquisition, such as inhomogeneous illumination [15]. Techniques such as morphological operations are applied to the input image. The following sessions elaborate the different preprocessing operations used in this paper.

1) Selection of Red Channel

In fundus images, the optic disc belongs to the brightest point of the image. The OD is often present in the red field as a well-defined white shape, brighter than the surrounding area. This color channel provides the highest contrast between OD and background [16]. So the red channel of the RGB colour image is used in this paper for the extraction of optic disc regions in the retinal fundus images.

2) Removal of blood vessels

Blood vessels within the OD act as strong distracters, and so they should be erased from the image beforehand. Mathematical morphology can extract important shape characteristics and also remove irrelevant information. It typically probes an image with a small shape or template known as a structuring element [15][17]. In this method a morphological closing operation is performed on the red channel. The erosion operation first removes the blood vessels and then the dilation approximately restores the boundaries to their former position. The closing operation is given in equation (1).

$$\text{Closing: } I_s(A, B) = A \bullet B = E(D(A, -B), -B) \quad (1),$$

where A is the red channel of the input image and B is a 10x10 symmetrical disc structuring element to remove the blood vessels, and IS represents the resultant vessel free, smoothed output image. Figure 1(a), Figure 1(b) and Figure 1(c) show the input image, the red channel of the input image, and the smoothed image. In the case of bright images intensity adjustment is also done.

B. Optic Disc Boundary Extraction

1) Fuzzy C Means Clustering combined with Thresholding

The smoothed image I_s is further processed for the extraction of the OD. The method presented here is a combination of Fuzzy algorithm, C Means clustering and Thresholding. Clustering involves the task of dividing data points into homogeneous classes or clusters so that items in the same class are as similar as possible and items in different classes are as dissimilar as possible. Most Fuzzy clustering algorithms are objective function based: They determine an optimal classification by minimizing an objective function. In objective function based clustering usually each cluster is represented by a cluster prototype. This prototype consists of a cluster center and maybe some additional information about the size and the shape of the cluster. The cluster center is an instantiation of the attributes used to describe the domain concerned. The size and shape parameters determine the extension of the cluster in different directions of the underlying domain. Depending on the data and the application, different types of similarity measures may be used to identify classes, where the similarity measure controls how the clusters are formed. In this new method intensity value is used as the similarity measure [18] [19]. Thresholding is one of the most powerful techniques for image segmentation in which the pixels are partitioned, depending on their intensity value, which would correspond to the background and the object [15]. The segmentation is then achieved by grouping all pixels having intensity greater than the threshold into one class, and all other pixels into another class.

2) Fuzzy C-Means Clustering Algorithm

Fuzzy C-Means (FCM) Clustering is a clustering technique which employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1.

It is an iterative algorithm. The aim of FCM is to find cluster centers (centroids) that minimize a dissimilarity function. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of difference between the cluster center and the data point. The more the data is near to the cluster center, the more is its membership towards the particular cluster center. Clearly, summation of the membership of each data point should be equal to 1. FCM is a method of clustering which allows one piece of data to belong to two or more clusters. It is based on minimization of the following objective function in equation (2).

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2, \quad 1 \leq m < \infty \quad (2)$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j , x_i is the i th of d -dimensional measured data, c_j is the d -dimension center of the cluster, and $\|*\|$ is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j given by:

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (3)$$

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m} \quad (4)$$

This iteration will stop when

$\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \epsilon$, where ϵ is a termination criterion between 0 and 1, whereas k -s are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m .

The algorithm is composed of the following steps:

- Step 1. Initialize $U = [u_{ij}]$ matrix, as $U(0)$
- Step 2. At k -step: calculate the centers vectors

$$C(k) = [C_j] \text{ with } U(k)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m}$$

Step 3. Update $U(k)$, $U(k+1)$

Step4. If $\|U(k+1) - U(k)\| < \epsilon$ then STOP; otherwise return to step2. [19][20]

The Fuzzy Logic Toolbox command line function, *fcm* is used for generating clusters in this paper. This function starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster. Next, *fcm* assigns every data point a membership grade for each cluster. In this paper three clusters are generated with low, medium and high membership grades. The outputs obtained are three cluster centers $C1$, $C2$ and $C3$ and membership function matrix M with membership-grades, which are the intensity values of pixels.

3) Thresholding

Thresholding is the operation of converting a multilevel image into a binary image i.e., it assigns the value of 0 (background) or 1 (objects or foreground) to each pixel of an image based on a comparison with some threshold value T (intensity or color value) [15]. As mentioned earlier the main feature of the OD is that it is having the highest intensity. The disc is extracted using the highest intensity and it is used as the threshold for the OD extraction. The threshold T is computed using the following method. From the generated clusters, first the cluster with maximum membership grade is found, and the corresponding grades are assigned with the same identification label. From the smoothed image, pixels with this gray level value are accessed, the average of the maximum and minimum intensity values are computed to obtain the threshold value T .

$$\text{i.e., } T = \frac{1}{2}[\text{Max}(\text{data}(\text{value})) + \text{Min}(\text{data}(\text{value}))] \quad (5)$$

In the above equation, data represents the data points of the smoothed image and label represents the cluster value with the highest membership grade. By applying the threshold T on the smoothed image I_s the image is converted to a binary image I_B . The following formula (6) [15] is used for the binary image extraction.

$$I_B(x, y) = \begin{cases} 1, & \text{if } I_s(x, y) > T \\ 0, & \text{if } I_s(x, y) \leq T \end{cases} \quad (6)$$

Another feature of the OD is that it is of circular shape. So the optic disc region selection process needs to be made specific to the circular region. So the largest connected component R_i whose shape is approximately circular is selected using the compactness measure

$$C(R_i) = \frac{P(R_i)}{4\pi A(R_i)} \quad (7),$$

where, $P(R_i)$ is the perimeter of the region R_i and $A(R_i)$ is the area of the region R_i . The binary image with the compactness smaller than the pre-specified value, (5 in the present study) is considered as the optic disc approximation. Thus using the condition $C < 5$, extraction of round objects is done, eliminating those objects that do not meet the criteria. In some cases the extracted image contains small unwanted objects. From the extracted optic discs, it is found that the area of optic disc is greater than 2500 pixels.

Hence, in order to remove the unwanted objects, the connected components or objects that have fewer than 2500

pixels are removed from the image producing the final binary image with the OD.

Fig. 2(a) shows the extracted binary image and fig. 2(b) shows the optic disc after removal of unwanted objects.



Fig. 2. (a) Extracted binary image (b) Extracted optic disc after removing small objects

From the extracted OD the exterior boundary is traced using the *bwboundaries* function from the tool box and this boundary is overlaid over the ground truth image. Figure 3(a), 3(b) and 3(c) manifest the ground truth boundaries obtained from two experts, shown in green and blue, together with the optic disc boundary obtained using the new method, in white colour overlaid on the ground truth image. It can be observed that the boundary traced by the ophthalmologists and the boundary estimated by the present method are close to each other.

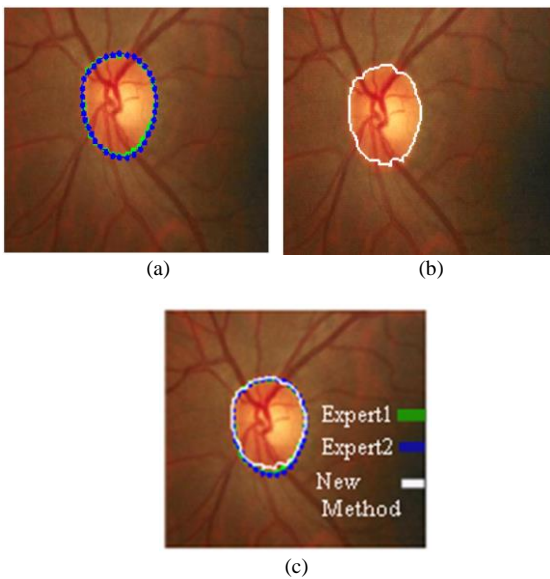


Fig. 3. (a) Ground truth boundaries (b) extracted boundary using the new algorithm (c) Overlay of (a) and (b)

IV. RESULTS AND DISCUSSIONS

A. Retinal Image Database

All the images used in this paper are obtained from the public database DRION-DB. It has 110 retinal images with each image having the resolution of 600 x 400 x 3 pixels and the optic disc annotated by two experts with 36 landmarks. The mean age of the patients was 53.0 years (standard Deviation 13.05), with 46.2% male and 53.8% female and all of them were Caucasian ethnicity. 23.1% patients had chronic simple glaucoma and

76.9% eye hypertension. The images were acquired with a colour analogical fundus camera, approximately centered on the ONH and they were stored in slide format. In order to have the images in digital format, they were digitized using a HP-PhotoSmart-S20 high-resolution scanner, RGB format, resolution 600 x 400 and 8 bits/pixel. Independent contours from 2 medical experts were collected by using a software tool provided for image annotation. In each image, each expert traced the contour by selecting the most significant papillary contour points and the annotation tool connected automatically adjacent points by a curve.

B. Implementation

The algorithm was applied on 110 images obtained from the DRION database. Ground truth is the OD boundary traced by 2 ophthalmologists. Five of the input image, along with their ground truth boundary as well as the detected boundary overlaid on the input image, is shown in fig.4 (a) and fig.4 (b) respectively.



Fig. 4. Examples for optic disc segmentation using the new algorithm (a) input image (b) ground truth and detected boundary overlaid

C. Performance Analysis

The performance analysis is done using the following parameters.

1) Mean Difference of Diameters(MDD)

The difference between both the vertical and horizontal diameter of the detected boundary and the corresponding diameter of the actual optic disc boundary are computed. Here the actual vertical and horizontal diameters are the ground truth boundary traced by the two experts. The three vertical diameters VD1, VD2 and VD3 and the three horizontal diameters HD1,

HD2 and HD3 are calculated using the boundaries from expert 1, expert 2 and the contour obtained using the new method, using of the equations (8) and (9).

$$VD_i = Y_{max_i} - Y_{min_i} \quad (8)$$

$$HD_i = X_{max_i} - X_{min_i} \quad (9)$$

where $i = 1, 2, 3$ and, X_{max_i} , X_{min_i} , Y_{max_i} , and Y_{min_i} are respectively the maximum and minimum coordinates of horizontal and vertical axes of the boundaries detected by expert 1, expert 2 and the present method.

Figure [5] shows the values of VD1, VD2 and VD3 as 2.9154, 2.9331 and 2.9369 and, HD1, HD2 and HD3 as 2.7671, 2.8796 and 2.9104 respectively, for a particular image. It can be seen that the 3 values of vertical diameters and 3 values of horizontal diameters are approximately the same.

Since, there was an inter-observer variability among the two experts tracing the contour of OD, therefore, the average of the contours traced by the two experts is used here. This work used as the gold standard, for each image, the average of the diameters obtained from the contours traced by the two experts

$$\text{i.e. } GV = \frac{1}{2}(VD1 + VD2) \quad (10)$$

$$GH = \frac{1}{2}(HD1 + HD2) \quad (11)$$

where GV and GH are the gold standards of vertical and horizontal ground truth diameters respectively. That is the average of the diameters obtained from the contours traced by the two experts.

The accuracy of the detected boundary is evaluated by the parameter MDD which is the mean of the differences of the gold standard diameters calculated using the equation (12).

$$MDD = \frac{1}{2} \left[\frac{1}{N} \sum_{i=1}^N (VD_i - GV) + \frac{1}{N} \sum_{i=1}^N (HD_i - GH) \right] \quad (12)$$

Fig. 5 shows the vertical diameter and horizontal diameter of an image with respect to expert 1, expert 2 and the new method.

As can be seen from the figure, the values of Table 1 shows the values of vertical and horizontal diameters of some of the images selected from the 110 fundus images used for the evaluation. It can be noted that the difference between the gold standard value and the value obtained by using the present method varies from 0 to 0.1298 in the case of vertical diameter and from 0 to 0.1190 in the case of horizontal diameter. The mean difference of the diameters (MDD) is evaluated as 0.0566. The smaller the MDD, the closer is the detected boundary to the ground truth.

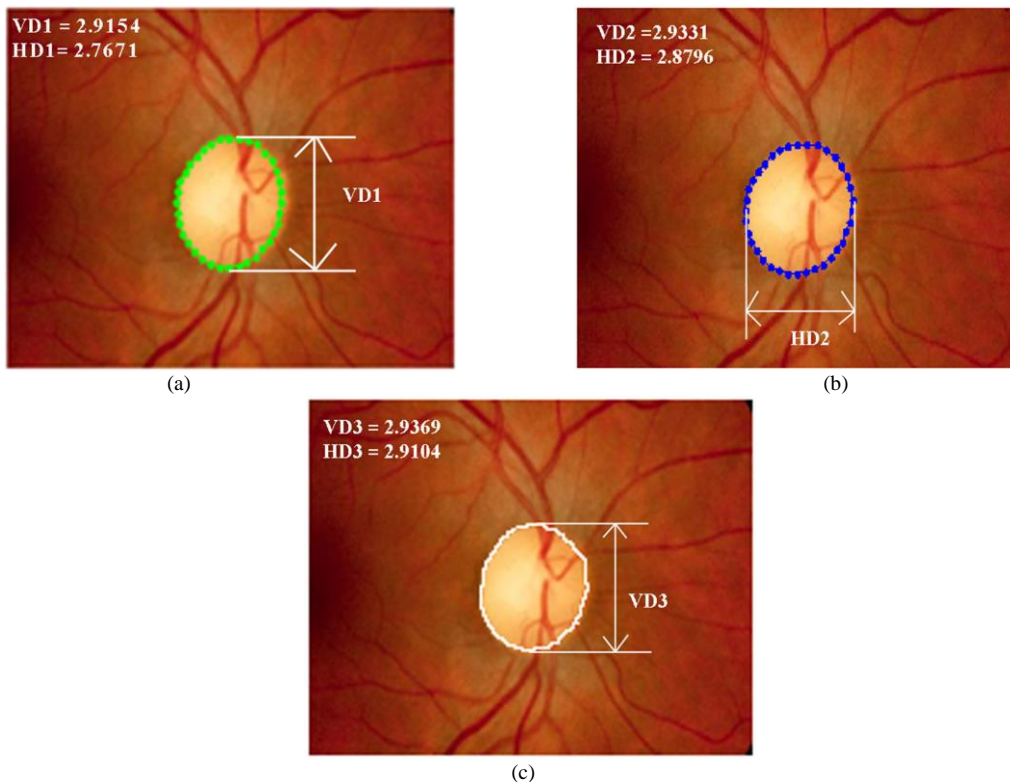


Fig. 5. The diameters obtained using the boundaries from (a) expert 1(VD1) (b) expert 2(HD2) and (c) present method (VD3)

TABLE I. EVALUATION OF HORIZONTAL AND VERTICAL DIAMETER OF A FEW FUNDUS IMAGES

| Image No. | Horizontal Diameter | | | | | Vertical Diameter | | | | | |
|--|---------------------|--------------|-----------------------|--------------------|----------------------|--------------------------------------|--------------|-----------------------|--------------------|---------------------|--------|
| | Expert 1 HD1 | Expert 2 HD2 | Gold standard GHD (1) | New Method HD3 (2) | Difference (1) - (2) | Expert 1 VD1 | Expert 2 VD2 | Gold standard GVD (3) | New Method VD3 (4) | Difference (3) -(4) | |
| Img1 | 2.6117 | 2.6413 | 2.6265 | 2.6194 | 0.0071 | 2.9026 | 2.9962 | 2.9494 | 2.9310 | 0.0184 | |
| Img2 | 2.3266 | 2.2269 | 2.2768 | 2.3548 | 0.0780 | 3.2015 | 3.2544 | 3.2280 | 3.2015 | 0.0265 | |
| Img3 | 2.3283 | 2.3548 | 2.3416 | 2.3813 | 0.0397 | 2.4077 | 2.4606 | 2.4341 | 2.5135 | 0.0794 | |
| Img4 | 2.6988 | 2.6988 | 2.6988 | 2.7840 | 0.0852 | 2.8575 | 2.8575 | 2.8575 | 2.9698 | 0.1123 | |
| Img5 | 2.5135 | 2.6194 | 2.5665 | 2.6194 | 0.0529 | 2.8310 | 2.8310 | 2.8310 | 2.7988 | 0.0322 | |
| Img6 | 2.4181 | 2.4671 | 2.4426 | 2.3283 | 0.1143 | 2.3330 | 2.3575 | 2.3452 | 2.3019 | 0.0433 | |
| Img7 | 2.6148 | 2.4575 | 2.5362 | 2.6488 | 0.1126 | 2.7052 | 2.7374 | 2.7213 | 2.7781 | 0.0568 | |
| Img8 | 2.2906 | 2.1952 | 2.2429 | 2.3019 | 0.0590 | 2.3456 | 2.2755 | 2.3106 | 2.1960 | 0.1146 | |
| Img9 | 3.0692 | 3.0956 | 3.0824 | 3.0692 | 0.0132 | 3.2015 | 3.2544 | 3.2280 | 3.2015 | 0.0265 | |
| Img10 | 2.2225 | 2.2225 | 2.2225 | 2.1167 | 0.1058 | 2.0902 | 2.0638 | 2.0770 | 2.0902 | 0.0132 | |
| Img11 | 2.4342 | 2.4342 | 2.4342 | 2.4342 | 0 | 2.3548 | 2.3548 | 2.3548 | 2.3813 | 0.0265 | |
| Img12 | 1.9844 | 2.0902 | 2.0373 | 2.0646 | 0.0273 | 1.9050 | 1.9315 | 1.9183 | 1.9696 | 0.0513 | |
| Img13 | 2.8575 | 2.8575 | 2.8575 | 2.9104 | 0.0529 | 2.7781 | 2.8046 | 2.7914 | 2.8369 | 0.0455 | |
| Img14 | 2.7586 | 2.6087 | 2.6837 | 2.7517 | 0.0680 | 2.4077 | 2.4871 | 2.4474 | 2.5723 | 0.1249 | |
| Img15 | 2.2907 | 2.2788 | 2.2848 | 2.2167 | 0.0681 | 1.9653 | 2.1624 | 2.0639 | 2.1225 | 0.0586 | |
| Img16 | 2.6988 | 2.7517 | 2.7253 | 2.8104 | 0.0851 | 2.5929 | 2.6194 | 2.6062 | 2.6867 | 0.0805 | |
| Img17 | 2.6269 | 2.7008 | 2.6639 | 2.6723 | 0.0084 | 2.6723 | 2.6723 | 2.6723 | 2.7781 | 0.1058 | |
| Img18 | 2.5400 | 2.5665 | 2.5533 | 2.5665 | 0.0132 | 2.6458 | 2.6458 | 2.6458 | 2.6723 | 0.0265 | |
| Img19 | 2.4606 | 2.4606 | 2.4606 | 2.4606 | 0 | 2.7252 | 2.6988 | 2.7120 | 2.7517 | 0.0397 | |
| Img20 | 3.1032 | 3.0153 | 3.0593 | 3.0245 | 0.0348 | 3.1426 | 3.2215 | 3.1821 | 3.1864 | 0.0044 | |
| Img21 | 2.6194 | 2.6458 | 2.6326 | 2.6723 | 0.0397 | 2.7781 | 2.5929 | 2.6855 | 2.5871 | 0.0984 | |
| Img22 | 2.4077 | 2.4077 | 2.4077 | 2.4606 | 0.0529 | 2.4077 | 2.1960 | 2.3018 | 2.3019 | 0 | |
| Img23 | 2.6964 | 2.7487 | 2.7225 | 2.8046 | 0.0821 | 2.7760 | 2.8910 | 2.8335 | 2.9633 | 0.1298 | |
| Img24 | 2.3283 | 2.2857 | 2.3070 | 2.2648 | 0.0422 | 2.6683 | 2.5585 | 2.6134 | 2.4871 | 0.1263 | |
| Img25 | 2.5929 | 2.6194 | 2.6062 | 2.7252 | 0.1190 | 2.2490 | 2.3019 | 2.2754 | 2.3019 | 0.0264 | |
| Mean difference of horizontal diameter | | | | | 0.0545 | Mean difference of vertical diameter | | | | | 0.0587 |
| Mean difference of diameters(MDD) | | | | | | | | | | 0.0566 | |

2) Karl Pearson Correlation

Karl Pearson Correlation gives the multiple correlations between the detected diameters and the ground truth diameters. Pearson Correlation is used to check the correlation between the obtained and detected diameters using the following formula.

$$r = \frac{N \sum xy - (\sum x)(\sum y)}{\sqrt{[N \sum x^2 - (\sum x)^2][N \sum y^2 - (\sum y)^2]}} \quad (13)$$

Where:

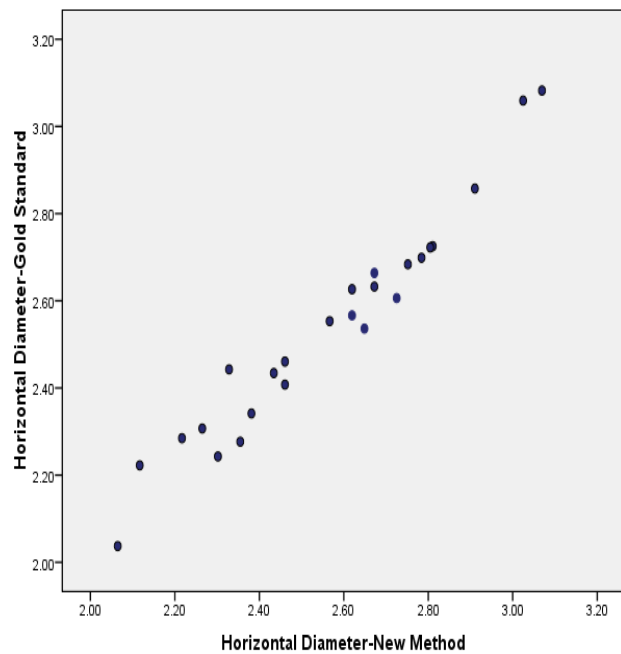
- N = number of pairs of scores
- $\sum xy$ = sum of the products of the paired scores
- $\sum x$ = sum of x scores
- $\sum y$ = sum of y scores
- $\sum x^2$ = sum of squared x scores
- $\sum y^2$ = sum of squared y scores

From the 25 test images, the Pearson Correlation of the obtained horizontal and vertical diameters to the ground truth diameters with respect to the two experts are obtained as 0.946 and 0.958 and 0.94 and 0.974 respectively.

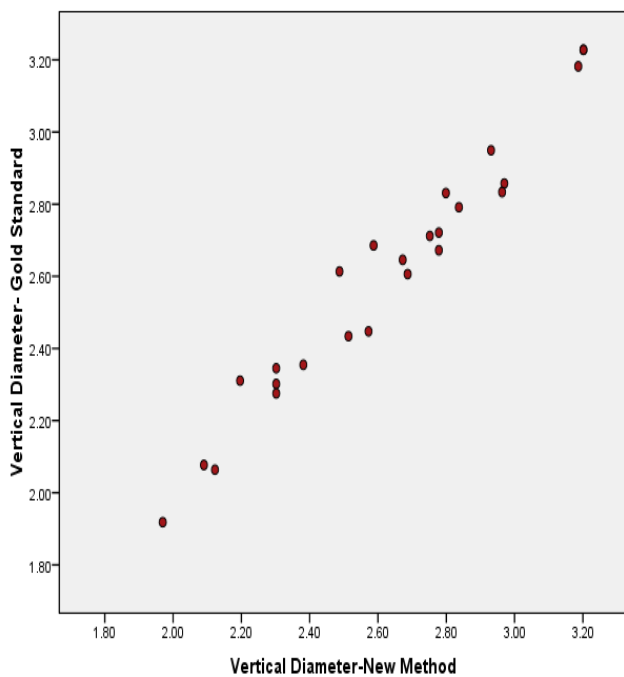
3) Scatter Plot Analysis

The scatter plot in Fig. 6(a) and Fig. 6(b) represents the detected vertical diameters Vs. the gold standard vertical diameters and detected horizontal diameters Vs. gold standard horizontal diameters with respect to the 25 test images.

It is clear from the scatter plot that there is a high positive linear correlation between the detected and the ground truth diameters.



(a)



(b)

Fig. 6. Scatter plot (a) detected vs. gold standard vertical diameter (b) detected vs. gold standard horizontal diameter.

V. CONCLUSION

A reliable and practical algorithm for the extraction of optic disc is developed in this paper. Fuzzy C Means algorithm combined with thresholding is used to extract the optic disc. The performance is evaluated using the proximity of the obtained disc contour with the ground truth contours from two experts. It is found that Fuzzy C means combined with thresholding provides a result which is close to the ground truth boundary. From the 25

test images, it is found that the Karl Pearson Correlation of the ground truth horizontal and vertical diameters from two experts with respect to the detected diameters are .946 & .958 and .94 & .974 respectively. The scatter plot depicts a high positive correlation between the gold standard and evaluated boundaries. The present method can achieve accurate segmentation for both normal and abnormal images. The main advantage of the development of this method is that it is fully automated and can be effectively used for the extraction of optic disc boundary of retinal images, and thus to save time for retinal image screening. At the same time, it is to be observed that, it does not extract optic discs accurately in certain bright images. However, as a future enhancement the advanced clustering technique could be applied to achieve more efficient accuracy in the result in the process of optic disc extraction.

REFERENCES

- [1] L. Gagnon, M. Lalonde, M. Beaulieu, M.-C. Boucher, L. Gagnon, M. Lalonde, M. Beaulieu, M.-C. Boucher, "Procedure to detect anatomical structures in optical fundus images Procedure to detect anatomical structures in optical fundus images", Computer Research Institute of Montreal; bDept. Of Ophthalmology, Maisonneuve-Rosemont Hospital.
- [2] A. D. Fleming, K. A. Goatman, S. Philip et al. "Automatic detection of retinal anatomy to assist diabetic retinopathy screening", Physics in Medicine and Biology 52, pp. 331–345, 2007.
- [3] Huiqi Li, OpasChutatape, "Automatic Location of Optic Disc in retinal Images", 0-7803-6725-1/01, IEEE, pp.837 – 840, 2001
- [4] Thitiporn Chanwimaluang and Guoliang Fan, "An efficient algorithm for extraction of anatomical structures in retinal images", Proc. of International Conference on Image Processing, Vol. 1, pp. 1093–1096, 2003
- [5] Hoover, A., Goldbaum, M., "Locating the optic nerve in a retinal image using the fuzzy convergence of bloodvessels", IEEE Transaction on Medical Imaging, Vol.22 Issue:8 pp.951 - 958 Aug. 2003
- [6] Chrastek R., Wolf M., Donath K., Niemann H., Paulus D., Hothorn T., Lausen B., Lammer R., Mardin C.Y., Michelson G. 'Automated Segmentation of the Optic Nerve Head for Diagnosis of Glaucoma', Medical Image Analysis, Vol.9, pp. 297-314, 2005.
- [7] Juan Xu, "Automated Optic Disc Boundary Detection by Modified Active Contour Model", IEEE Transactions on Biomedical Engineering", Vol.54, No.3, pp.473-482, 2007
- [8] Enrique J. Carmona, Mariano Rinco'n , Julia'n Garc'a- , Jose´ M. Mart'nez-de-la-Casa , "Identification of the optic nerve head with genetic algorithms", Artificial Intelligence in Medicine (2008) 43, 243–259, Elsevier 2008
- [9] Echegaray, S; Soliz, P; Luo, W , "Automatic initialization of level set segmentation for application to optic disc margin identification", CBMS 2009. 22nd IEEE International Symposium on Computer-Based Medical System, Proc. pp. 1 – 4, 2009.
- [10] Welfer D, Scharcanski J, Kitamura CM, Pizzol MMD, Ludwig LWB, Marinho DR, " Segmentation of the Optic Disk in Color Eye Fundus Images Using an Adaptive Morphological Approach", Computers in Biology and Medicine, 40(2): 124-137, 2010.
- [11] P. C. Siddalingaswamy and P. K. Gopalakrishna, "Automatic Localization and Boundary Detection of Optic Disc Using Implicit Active Contours", International Journal of Computer Applications, vol. 1, no. 6, pp. 1-5, 2010.
- [12] Yuji Hatanaka, Atsushi Noudo, Chisako Muramatsu, Akira Sawada, Takeshi Hara, Tetsuya Yamamoto, and Hiroshi Fujita, "Automatic Measurement of Cup to Disc Ratio Based on Line Profile Analysis in Retinal Images", 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, pp.3387 – 3390, August 30 - September 3, 2011
- [13] Amin Dehghani, Hamid Abrishami Moghaddam and Mohammad-Shahram Moin, "Optic disc localization in retinal images using histogram matching", EURASIP Journal on Image and Video Processing 2012, 2012:19 doi:10.1186/1687-5281-2012-19, 2012.
- [14] Angel Suero, Diego Marin, Manuel E. Gegundez-Arias, and Jose M.

- Bravo, "Locating the Optic Disc in Retinal Images Using Morphological Techniques", IWBBIO 2013. Proceedings, Granada, 18-20 March, 2013, pp.593-600, 2013.
- [15] Rafael C Gonzalez, Richard E Woods, Steven L Eddins, Digital Image Processing, Prentice Hall Publications, 2008.
- [16] T. Walter and J. C. Klein, "Automatic analysis of colour fundus photographs and its application to the diagnosis of diabetic retinopathy," in Handbook of Biomedical Image Analysis. New York: Kluwer, vol. 2, pp. 315-368 2005.
- [17] Rafael C Gonzalez, Richard E Woods, Steven L Eddins., Digital Image Processing Using Matlab, Prentice Hall Publications, 2004.
- [18] Yinghua Lu, Tinghui Ma, Changhong Yin, Xiaoyu Xie, Wei Tian, ShuiMing Zhong, "Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data", International Journal of Database Theory and Application, Vol.6, No.6, pp.1-18, 2013.
- [19] HeikoTimm, Christian Borgelt, and Rudolf KruseFuzzy, "Cluster Analysis with Cluster Repulsion", CiteSeerx.
- [20] Yong Yang , Shuying Huang, "Image segmentation by fuzzy c-means Clustering algorithm with a novel Penalty term", Computing and Informatics, Vol. 26, pp. 17-31, 2007.

Feature Descriptor Based on Normalized Corners and Moment Invariant for Panoramic Scene Generation

Kawther Abbas Sallal
Computer Science Department
Babylon University
Babylon, Iraq

Abdul-Monem Saleh Rahma
Computer Science Department
University of Technology
Baghdad, Iraq

Abstract—Panorama generation systems aim at creating a wide-view image by aligning and stitching a sequence of images. The technology is extensively used in many fields such as virtual reality, medical image analysis, and geological engineering. This research is concerned with combining multiple images with a region of overlap to produce a wide field of view by the detection of feature points for images with different camera motion in an efficient and fast way. Feature extraction and description are important and critical steps in panorama construction. This study presents techniques of corner detection, moment invariant and random sampling to locate the important features and built storing descriptors in the images under noise, transformation, lighting, little viewpoint changes, blurring and compression circumstances. Corner detection and normalization are used to extract features in the image, while the descriptors are built by moment invariant in an efficient way. Finally, the matching and motion estimation is implemented based on the random sampling method. The results of experiments conducted on images and video sequences taken by handheld camera and images taken from the internet. The results show that the proposed algorithm generates panoramic image and panoramic video of good quality in a fast and efficient way.

Keywords—Feature extraction; feature description; motion estimation; registration; panoramic scene

I. INTRODUCTION

Panoramic view construction is one of the most computer vision applications that have a great attention recently. The technology of panoramic view is developed rapidly and becomes a kind of popular visual technology, because the visual panorama technology can bring people a new real visualization of the scene and interactive experience [1][2]. It aims at creating a wide view image by aligning and stitching a sequence of images that having a significant overlap. The technology is extensively used in many fields such as virtual reality, medical image analysis, mapping, visualization, and geological engineering [2].

Image registration operation is very important for panoramic view generation. Image Registration is the process of matching two or more images of the scene. This requires the estimation geometric transformations to align the images with respect to a common reference. Image registration is of great importance in all processing and analysis tasks based on the combination of data from sets of images. Image registration algorithms can be divided into two major categories: feature-based methods and area-based methods.

Feature-based methods find relevant image features, known as control points, such as corners, point-like structures, line intersections, line ending points or high-curvature points that can be matched between two or more images. Once a sufficient number of points have been matched by correspondence on two images, a suitable geometric transformation can be computed and applied to align them.

Area-based methods, also known as correlation based or template matching methods, work by finding correspondences between regions of the images without considering any features. Some of these algorithms are based on cross correlation in the spatial or frequency domain, or on mutual information. The correlation can be estimated locally, for example, for squared regions distributed over a regular lattice, or globally for the whole image. If two images are correlated, then the registration process continues by finding the parameters of a geometric transformation that maximizes cross correlation, and the images are aligned accordingly.

The image registration process is one of the most complex and challenging problems of image analysis, where the extreme diversity of images and working scenarios make impossible for any registration algorithm to be suitable for all applications.

II. RELATED WORK

Many methods have been presented in recent years. Szeliski in [3] looks at one way to use video as a new source of high-resolution. Video is a low-resolution medium that compares poorly with computer displays and scanned imagery. It also suffers, as do all input imaging devices, from a limited field of view. He present algorithms that align frames of video and composite scenes of increasing complexity beginning with simple planar scenes and progressing to panoramic scenes and, finally, to scenes with depth variation. His approach directly minimizes the discrepancy in intensities between pairs of images after applying the recovered transformation.

In [4] the researchers used an algorithm based on Correlation. For Automatic Image Registration Applications, the features like edges are detected by using Sobel Edge Detection Algorithm. For matching the features, first Segmenting the image file in terms of different blocks and then applying the Hierarchical matching to create a pyramid of blocks. Finally, applying correlation based matching starting from the top level of the pyramid. Otherwise take a suitable

pixel block size say about 32 x 32 pixel block from right image and search for the exact location of that 32 x 32 pixel block in the left image. [5] Proposed the video serial images registration based on a feature based method algorithm. A matching method based on the improved algorithm can get better image fusion and image registration which is introduced for attaining more precise aggregate of matching points. In order to estimate the fundamental matrix which describes the whole geometry accurately and robustly, an improved SVD decomposition with weighted normalized fundamental matrix calculating method is proposed. [6] Describes a simple method for taking two videos and creating a panoramic video. Because using of hand held camera, it is difficult to obtain the case of stability. Therefore to get the optimization case, the researchers assume there is no change in acquiring circumstances of video. They assumed that the transformation between the two frames is the translation case. The translation between the two frames is estimated based on the matching of moment values for selected points in the overlapping regions. Also, they assume to use motion estimation techniques like three step search for the stitched frames to produce a compressed panoramic view. The limitation of this algorithm is, it works with the optimal case because this method is not excellent when the distorted transformations occurred or the gray values difference between the two images is found. [7] Presents an approach for the panoramic view generation. First, salient features are robustly detected from the input images by a robust algorithm called Scale Invariant Feature Transform (SIFT). SIFT features are invariant to translation, rotation, image scaling and partially invariant to viewpoint, illumination changes and image noise. These features are matched between the successive images and hence image transformation is estimated. Then, the image blending technique blends the images together to get a panoramic view without visible edge seam. [8] Proposes the feature based image fusion approach. The fusion image system includes features point detection, feature point descriptor extraction and matching. A RANSAC algorithm is applied to eliminate the number of mismatches and obtain a transformation matrix between the images. The input image is transformed with the correct mapping model for image stitching. In this paper, feature points are detected using steerable filters and Harris, and compared with traditional Harris, KLT, and FAST corner detectors.

Our contribution is introducing a new direction for developing the used multimedia and devising new Media more arousing and attracting for viewers. An efficient method for invariant feature detection is introduced based on mixing local and global feature extraction methods. Some researchers worked on designing the panoramic image, but this work aims to develop an algorithm to produce the panorama in a simple and accurate way and making it as a seed in generating films of that type. The remainder of this paper is organized as follows. Section 2 describes the general system. Section 3 includes the methodology. Section 4 explains the steps for features detection and descriptors construction. Section 5 explains the feature matching operation and motion estimation. Section 6 discusses the experimental results and section 7 explains the conclusions.

III. METHODOLOGY

The panoramic scene is constructed by alignment of multiple images with overlapping region. The alignment operation requires detection of similar regions in images. The similarities between images are determined by an efficient feature extraction method. The first step is capturing two or more images using handheld camera. The second step is enhancing the image details by using median filter. Then, the important features in each image are extracted using a local feature extraction method. The extracted features are normalized. Normalization is important step if the illumination invariance is required. Each extracted feature needs to calculate a descriptor that determines its invariability under different circumstances. The feature descriptor is calculated using a famous region feature extraction method called moment invariant. After that, the feature descriptors for both images are matched using a distance metric. The matching result is refined using second matching step based on random sampling method and the motion between the two images is estimated accordingly. The final step is using the estimated motion in the construction of the panoramic scene. The complete steps of the proposed system are shown in fig. 1 below. A hand held 4300S Nikon camera with a 16 mega-pixel resolution is used. Since the images and video sequences are taken by the user. The captured images and video sequences must have region of overlap to satisfy the condition of generating the panoramic view.

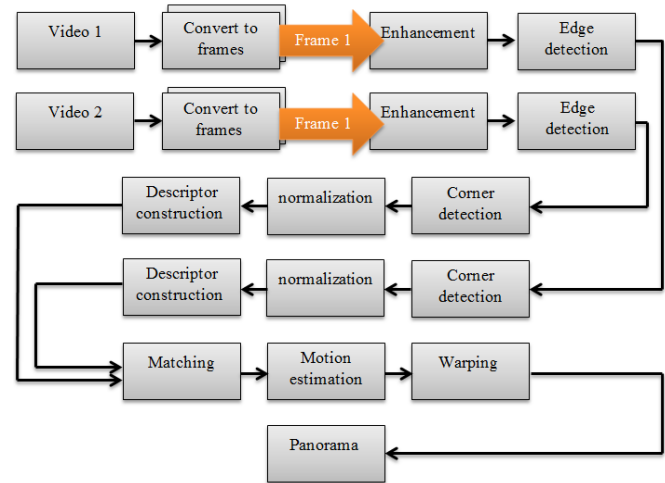


Fig. 1. The proposed system

IV. IMAGE PREPROCESSING

It is desirable to perform some kind of noise reduction on an image before the analysis process. Because the edge detection will be used later in feature extraction process, it is important to enhance the image details first. The median filter is a nonlinear digital filtering technique, often used to remove noise and used here to enhance the edge details. Median filtering is very widely used in digital image processing because, under certain conditions, it preserves edges while removing noise. The main idea of the median filter is to run through the image pixel by pixel, replacing each pixel with the median of neighboring pixels. The pattern of neighbors is called the "window", which slides, pixel by pixel,

over the entire image. Fig. 2 shows the result of applying median filter with window 3 x 3. In this work, the image is filtered with median filter first, then the filtered image is subtracted from the original image to obtain the details image. Then, the details image is summed with the filtered image to get the final filtered image.

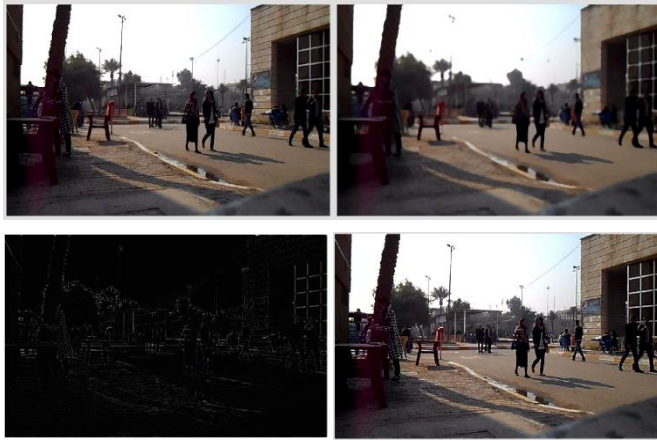


Fig. 2. Implementation of the details enhancement method

V. THE IMPROVED CORNER DETECTOR

The harris autocorrelation detector [9] is a development detector of the moravec's detector. The 'corner' is a location in the image where the local autocorrelation function has a distinct peak. Corner point detection has found its application in various computer vision tasks. In this work, improved corner detector method is proposed to extract corner information as the first step of the proposed algorithm.

This method not only solved the problem of the discrete shifts, but also it deals with the issue of directions with the advantage of the autocorrelation function and increased the accuracy of localization. Feature point extract by the Harris operator has a rotation and translation invariability and has a good robustness against noise and change of parameters during acquisition of image. The improved method consists of the following steps:

1) Enhance the input image details using median filter. Then convert the RGB image to gray color image.

2) To get rid of the noise, the image is smoothed with Gaussian filter with sigma σ . larger σ increasing the smoothing.

$$G_{x,y} = \frac{1}{2\pi\sigma^2} \exp^{-(x^2+y^2)/2\sigma^2} \quad (1)$$

Where x is the distance from the origin in the horizontal axis, y is the distance from the origin in the vertical axis, and σ is the standard deviation of the Gaussian distribution.

3) Apply prewitt edge detection algorithm on the smoothed image by convolving two masks, horizontally and vertically to obtain the first derivatives in x -direction (f_x) and y -direction (f_y) as following:

| | | |
|--------------|------------|--------------|
| $f(i-1,j-1)$ | $f(i,j-1)$ | $f(i+1,j-1)$ |
| $f(i-1,j)$ | $f(i,j)$ | $f(i+1,j)$ |
| $f(i-1,j+1)$ | $f(i,j+1)$ | $f(i+1,j+1)$ |

Figure (4-13) pixel neighbors

The observation pixel is (i,j) .

$$f_x = f(i+1,j-1) + f(i+1,j) + f(i+1,j+1) - f(i-1,j-1) - f(i-1,j) - f(i-1,j+1) \quad (2)$$

$$f_y = f(i+1,j+1) + f(i,j+1) + f(i-1,j+1) - f(i-1,j-1) - f(i,j-1) - f(i+1,j-1) \quad (3)$$

4) Three values must be obtained from the result above. These values are called the second order moment. The first value is the square of gradient in x -direction (f_x^2). The second value is the square of gradient in y -direction (f_y^2). The third value is the multiplication of gradient in x -direction and gradient in y -direction ($f_x f_y$).

5) The resulted images from the previous step is smoothed again to keep as possible only the true corners.

6) Compute the corner value (R) using the following equation:

$$f(x, y, \delta 1) = f(x, y) * G(\delta 1) \quad (4)$$

$$R = \frac{G(\delta 2) f_x^2(x, y, \delta 1) * G(\delta 2) f_y^2(x, y, \delta 1) - (G(\delta 2) f_x f_y(x, y, \delta 1))^2}{G(\delta 2) f_x^2(x, y, \delta 1) + G(\delta 2) f_y^2(x, y, \delta 1) + \beta} \quad (5)$$

Where δ_1 and δ_2 is the standard deviation of the Gaussian distribution (G), $f_x(x,y)$ and $f_y(x,y)$ are the partial derivatives of $f(x,y)$ in x and y directions, and β is a constant value.

7) Apply the non-maxima suppression to extract the maximum value within predefined neighbors. The non-maxima suppression is like a filter which only lets the value pass if it is the maximum of its neighbors.

$$R_i > R_j \quad \forall j \in N_i \quad (6)$$

Where R_i and R_j are the corner values of pixel i and its neighbor pixel j . N_i is the N neighbors of pixel i in a predefined window.

8) Compute the mean and standard deviation for the result $ME = \sum_r \sum_c \frac{R(r,c)}{N}$ (7)

$$std = \sqrt{\frac{1}{N} \sum (R_{(r,c)} - ME)^2} \quad (8)$$

9) Obtain the strongest or the more stability corners by normalization of the corners using the following equation:

$$NC(r,c) = R(r,c) - ME / std \quad \geq T \quad (9)$$

Where, NC is the normalized corner, R is the corner response value, ME is the mean value, std is the standard deviation value and T is the threshold.

10) Extract the final corners if the result of the above equation is above the predefined threshold, else ignore the corner. The result is shown in fig.3.

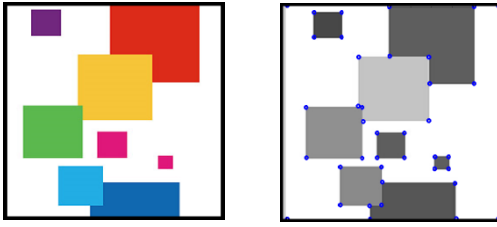


Fig. 3. corner feature extraction

VI. DESCRIPTOR CONSTRUCTION

One of the main challenges lies in classifying and recognizing objects from different views and lighting conditions. Depictions of natural scenes typically do not maintain their viewpoint, having rotational, perspective, projective and zoom changes between images of the same object. Interest points have to focus on the same locations of an object, no matter from which point of view they are shown. When provided with stable interest points under these circumstances, local descriptors become more effective than using random features of an object. In the previous step, the important features are gotten but it is necessary to identify each feature point. Therefore, an efficient descriptor will be generated for each detected feature point. For each feature, an $N \times N$ window centered on this feature is determined. Then, the window is divided to $n \times n$ blocks where $N > n$. For each block the average of geometrical moments is calculated. The result is a descriptor of $n \times n$ matrix for each feature point.

Moment invariants are the most popular and widely used shape descriptors in computer vision derived by Hu. A 2-D function $f(x, y)$ of the order $(p + q)$ is defined as [10]:

$$\mu_{pq} = \sum_x \sum_y (x)^p (y)^q f(x, y) \quad (10)$$

For $p, q = 0, 1, 2, \dots$

The uniqueness theorem states that if $f(x, y)$ is piecewise continuous and has non zero values only in a finite part of x - y plane, moments of all orders exist and the moment sequence (m_{pq}) is uniquely determined by $f(x, y)$. Conversely, (m_{pq}) uniquely determines $f(x, y)$. The central moments can be expressed as [10][11][12]:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x, y) \quad (11)$$

$$\text{Where } \bar{x} = \frac{m_{10}}{m_{00}} \text{ and } \bar{y} = \frac{m_{01}}{m_{00}}$$

$$\mu_{00} = m_{00}, \mu_{10} = 0, \mu_{01} = 0$$

The normalized central moments, denoted η_{pq} , are defined as

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^\gamma} \quad (12)$$

$$\text{Where } \gamma = \frac{p+q}{2} + 1$$

A set of seven invariant moments can be derived from the second and third moments:

$$\phi_1 = \eta_{20} + \eta_{02} \quad (13)$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \quad (14)$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \quad (15)$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \quad (16)$$

$$\begin{aligned} \phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ &+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{aligned} \quad (17)$$

$$\begin{aligned} \phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ &+ 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \end{aligned} \quad (18)$$

$$\begin{aligned} \phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ &+ (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{aligned} \quad (19)$$

The seven invariant moments, which are invariant to translation, scaling, mirroring and rotation, composed of the linear combination of the second-order and third-order central moments. Because of the seven moment invariants is relatively large, and to simplify comparison, making use of logarithmic methods. At the same time, taking into account the possible negative moment invariants situation, you have to take the absolute value before getting logarithm.

VII. MOTION ESTIMATION FROM CORRESPONDENCES

Images can be in different transformations that can be resulted during camera acquiring. According the complexity, these transformations are rigid, affine, non-rigid. Rigid registration models are linear and only allow for translation, rotation and scale changes without any distortion. Affine transform is also linear and support overall distortions besides shearing and stretching. Non-rigid models are nonlinear and allow for arbitrary local and global distortions. These transformations effect on matching operation. Therefore, it is important to find a way to determine the true matches. The true matches can be determined if the points fit with a predefined model. The matching is done by computing the Euclidian distance between two descriptors depending on the second nearest neighbor technique. The matching result contains a number of error matches which effect on the motion estimation results. Therefore, a random sampling method is used to get the invariance matches and estimating the motion between the images.

1) Nearest Neighbor Matching

The Nearest Neighbor algorithm uses the ratio of distance between nearest neighbor feature points to that of second nearest neighbor feature points to match feature points. Using the ratio of nearest neighbor to second nearest neighbor to match feature points can obtain a good result, because a correct registration will have a more obvious shortest distance of nearest neighbor than that of misregistration, which will achieve a stable registration. Assume (i) as the feature point in image 1, and (j) as the feature point of nearest neighbor in image 2. If the ratio of the nearest distance to second-nearest distance (a) is less than a certain threshold, then this pair of registration points is matched, as described below:

$$R = D(i, j) / D(i, a), \quad (20)$$

If $R < T$ then (i and j) is matched

Where $D(i, j)$ is the distance between point (i) in the first image and point (j) in the second image and (a) is the second nearest neighbor point as the following equation:

$$D(i, j) = \sqrt{(i - j)^2} \quad (21)$$

The points (i) and (j) is matched if the value of R is lower than the predefined threshold T. If T is decreased, the number of registration points will be reduced but more stable. The mismatched points can be regarded as outliers, which are the data that do not conform to the model. These outliers can rigorously disturb the estimated motion, and consequently should be identified.

2) Random Sampling Technique

It's important to determine a set of invariance matches, which are the data whose distribution can be explained by some set of model parameters from the presented correspondences so that the transformation can be estimated in an optimal manner. In the computer vision field, any two images of the same planar surfaces are related by a transformation. This is very important in computing the camera movement, like rotation and translation and other transformation between two images. In mathematical definition the homogeneous coordinates are used, because matrix multiplication cannot be used directly to perform the division required by the perspective projection.

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \quad (22)$$

Where $p' = Tp$.

Each point correspondence generates two linear equations for the elements of T (dividing by the third component to remove the unknown scale factor).

$$x' = \frac{t_{11}x + t_{12}y + t_{13}}{t_{31}x + t_{32}y + t_{33}} \quad (23)$$

$$y' = \frac{t_{21}x + t_{22}y + t_{23}}{t_{31}x + t_{32}y + t_{33}} \quad (24)$$

And multiplying out

$$x'(t_{31}x + t_{32}y + t_{33}) = t_{11}x + t_{12}y + t_{13} \quad (25)$$

$$y'(t_{31}x + t_{32}y + t_{33}) = t_{21}x + t_{22}y + t_{23} \quad (26)$$

Then, $n \geq 4$ points generate $2n$ linear equations, which are sufficient to solve T. The above equation can be rearranged as:

$$\begin{pmatrix} x & y & 1 & 0 & 0 & 0 & -x'x & -x'y & -x' \\ 0 & 0 & 0 & x & y & 1 & -y'x & -y'y & -y' \end{pmatrix} T = 0 \quad (27)$$

Where,

$T = (t_{11}, t_{12}, t_{13}, t_{21}, t_{22}, t_{23}, t_{31}, t_{32}, t_{33})^T$ is the matrix T written as a vector.

The random sampling algorithm [13] is suggested in this work to be applied on the initial matches to determine the invariance matches. This algorithm was first introduced by Fischler and Bolles as a method to estimate the model's parameters in the presence of large amounts of variance matches. It has been widely used in the computer vision and image processing for many different purposes.

This algorithm includes two steps that are repeated in an iterative fashion. First, a set of points is randomly selected from the input dataset and the model parameters are computed using only the elements of this set as opposed to least squares, where the parameters are estimated using all the data available. In the second step the algorithm checks which elements of the full data set are consistent with the model instantiated with the parameters estimated in the first step. The algorithm ends when the probability of finding better set points is below a certain threshold [14]. In this work four points are randomly selected from the set of candidate matches to compute the transformation. Then, select all the pairs which agree to the transformation. A pair (p; p') is considered to agree to a T, if:

$$\text{Dist}(T, p; p') < \epsilon, \quad (28)$$

For some threshold ϵ (represents the amount of error) and Dist is the Euclidean distance between two points. The third step is repeating the previous two steps until enough pairs are consistent with the computed transformation. The results from this step are the group of invariance matching features and the motion matrix that the second image is wrapped accordingly to generate the panoramic view.

VIII. GAUSSIAN MASK COLOR ADJUSTING

Due to various reasons, including the light, the geometry of the camera and other reasons, the overlapping regions of frames are almost never the same. When mosaicking the two frames, the resulted frame contains a distinctive seam. A seam is the artificial edge produced by the intensity differences of pixels immediately next to where the images are joined. Therefore, to avoid the intensity disparity on the mosaicking line, a color adjusting step must be done. To adjust the color of the resulted frames, the frames are combined with a background (using Gaussian function) to create the appearance of partial transparency. It is often useful to render image elements in separate passes, and then combine the resulting multiple frames into a single frame, final frame in this process called the composite frame.

The first step in color adjusting process is determining the suitable distance in image for blending and this depends on the computed motion. Then, the distance value is used to build two masks for the images. The masks are filtered with Gaussian function to introduce the transparent background. Then, each image will be multiplied with its mask to produce the final color adjusted image.

IX. EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate the performance of the proposed algorithm, the experiments include images and video sequences taken by handheld camera in different situations. For each image, after details enhancement with median filter, the initial corners and the normalized corners are extracted depending on its mean and standard deviation and shown in home image in fig. 4. The value of the used threshold must be sufficient to extract enough number of corners. After extracting the final corner points in the two images, the descriptors for them are created by taking a 25*25 window around each point, divide the window to blocks of 5*5 each and finding the average value of moment invariant for them.

Then, the matching operation is done by finding the nearest neighbor. A very important factor here is the matching threshold because it must be chosen to get as more matching points as possible. The result of this step illustrated in fig. 5. The next step is eliminating the mismatch points to obtain only the invariance matches that represent the key feature points by using a random sampling algorithm.

The number of iterations is an important factor here to obtain more matches between the images. The result of this step is shown in fig. 6. Finally, the panoramic scene is constructed according to the estimated motion from the previous step. The projected image is shown in fig. 7.

The constructed panorama before and after color adjusting is explained in fig. 8. Second test of the proposed method are shown in fig. 9. Fig.10 explains an example on panoramic video generation using the proposed method. For example, five frames are taken from each video.



Fig. 4. The detected and the normalized corners



Fig. 5. The initial matches



Fig. 6. The invariance matches

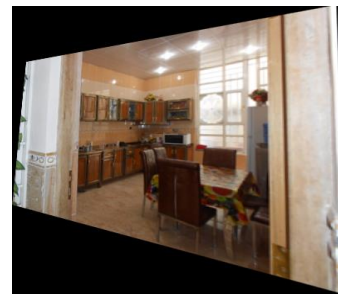


Fig. 7. Transformed image according the estimated T

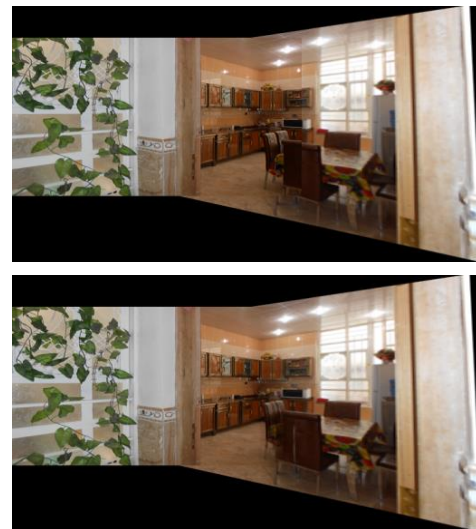


Fig. 8. Panoramic image before and after color adjusting



Fig. 9. Panoramic image from ice age movie



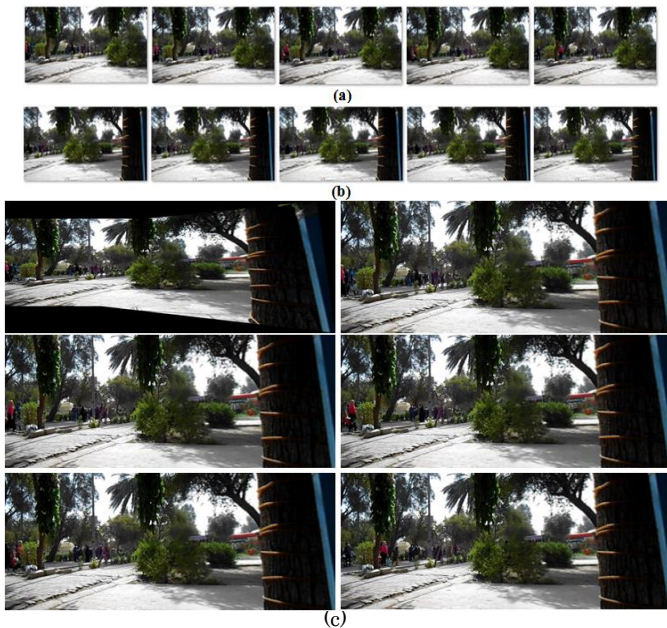


Fig. 10. Panoramic video for Al_Zawraa city. frames(1,9,19,26,100) in: (a) video 1 and (b) video 2 and (c) panoramic frames.

As shown in the figures and in experiments, the accuracy of the proposed algorithm is coming from the using of the random sampling algorithm because it follows the object's motion to estimate the matched feature points under different circumstances like transformations, lighten, noise, view point. For, panoramic video generation, the proposed method was applied on the first frame only. The other frames will be accompanied according to the estimated motion directly for fast execution.

X. CONCLUSIONS

Panoramic scene generation is an important topic because it needs a fusion of image processing, graphics and computer vision techniques. Many researchers deal with generating panoramic image depending on the SIFT method but in this work, a panoramic image using a proposed feature extraction method is generated. The extracted features and constructed descriptors in the overlapping region of the images are based on corner points, geometrical moments and random sampling for feature matching and motion estimation from initial matches. The mixing of local feature extraction method represented by improved corner detector and region feature extraction method represented by geometrical moments and random sampling for feature filtering and motion estimation

gives us a fast, efficient, accurate method and less complexity than the famous SIFT method. Also, the proposed method is applied on two video sequences and gives very good result according to many factors like in time and in quality of the resulted panoramic video. The execution time changed due to the number of extracted features and the number of iterations used in random sampling method. For images the execution time between 6-35 sec. while for videos it is between 380-260 sec. the quality of image before and after color adjusting is measured by PSNR and for all images and video sequence samples between 33-43dbi.

REFERENCES

- [1] E. Zheng, R. Raguram, and P. F. George, "Efficient generation of multi-perspective panoramas", IEEE, International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, 2011.
- [2] Yun Liu, Jiru Zhang, Jian Gao, "Panoramic Technique in the Video Monitoring System and Implementation", TELKOMNIKA, Vol.11, No.1, pp. 91-96, ISSN: 2087-278X, 2013.
- [3] R. Szeliski, "Video mosaics for virtual environments", IEEE Computer Graphics and Applications, 1996, pp: 22-30.
- [4] R. Asha, K. Asha and M. Manjunath, "Image Mosaicing and Registration", International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013.
- [5] W. Bing and W. Xiaoli, "Video Serial Images Registration Based on FBM Algorithm", Research Journal of Applied Sciences, Engineering and Technology 5(17), 2013.
- [6] A. Kawther and S. Abdul Monem, "Generation of Video Panorama System", International Journal of Computer Applications, USA vol.73 issue 5, pp: 20-26, July 2013.
- [7] L. Lin and G. Nan, "Algorithm for Sequence Image Automatic Mosaic based on SIFT Feature", WASE International Conference on Information Engineering, 2010.
- [8] M. V. Subramanyam and Mahesh, "Feature Based Image Mosaic Using Steerable Filters and Harris Corner Detector", I.J. Image, Graphics and Signal Processing, 6, 2013, pp:9-15.
- [9] C. Harris and M. Stephens, "A combined corner and edge Detector", In Alvey Vision Conference. 1988, Pp. 147-151.
- [10] M.K. Hu, "Visual pattern recognition by moments Invariants", IRE Trans. Information Theory, Vol 8. 1962, Pp.: 179- 87.
- [11] R. Mohamed, Y. Haniza, S. Puteh, Y. Ali, S. Abdul R., R. Mohd, Y. Sazali, D. Hazri and M. Karthigayan, "Object Detection using Geometric Invariant Moment", American Journal of Applied Sciences. Vol 3 no.6. 2006, Pp.: 1876-1878.
- [12] J. Bei and L.Chen, "Map Matching Algorithms Based on Invariant Moments", Journal of Computational Information Systems Vol. 7 (16). 2011, Pp.: 5668-5673.
- [13] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography", Communications of the ACM, vol. 24. No. 6. 1981, Pp.: 381-395.
- [14] M. Zuliani, "Computational Methods for Automatic Image Registration", PHD thesis. University of California, USA, 2006.

Hybrid Client Side Phishing Websites Detection Approach

Firdous Kausar, Bushra Al-Otaibi, Asma Al-Qadi, Nwayer Al-Dossari

Department of Computer Science
Imam University
Riyadh, Saudi Arabia

Abstract—Phishing tricks to steal personal or credential information by entering victims into a forged website similar to the original site, and urging them to enter their information believing that this site is legitimate. The number of internet users who are becoming victims of phishing attacks is increasing beside that phishing attacks have become more sophisticated. In this paper we propose a client-side solution to protect against phishing attacks which is a Firefox extension integrated as a toolbar that is responsible for checking whether recipient website is trusted or not by inspecting URLs of each requested webpage. If the site is suspicious the toolbar is going to block it. Every URL is evaluated corresponding to features extracted from it. Three heuristics (primary domain, sub domain, and path) and Naïve Bayes classification using four lexical features combined with page ranking received from two different services (Alexa, and Google page rank) used to classify URL. The proposed method requires no server changes and will prevent internet users from fraudulent sites especially from phishing attacks based on deceptive URLs. Experimental results show that our approach can achieve 48% accuracy ratio using a test set of 246 URL, and 87.5% accuracy ratio by excluding NB addition tested over 162 URL.

Keywords—Phishing Attacks; Browser Plugin; Anti Phishing; Security; Firefox

I. INTRODUCTION

Phishing is an online identity theft in which attackers use social engineering to appear as a trusted identity to gain valuable information. Phishing exploits human vulnerabilities rather than software vulnerabilities. It targets many kinds of confidential information including usernames, passwords, social security numbers, credit card numbers, bank account, and other useful personal information.

In the past few years we have seen an increase in the number of phishing attacks with many variants of techniques targeting every sector of society. As reported by the Anti-Phishing Working Group (APWG) (Anti-Phishing Working Group. "Phishing Activity Trends Report: Third Quarter 2013 Report, 2014) "Payment Services continued to be the most-targeted industry sector throughout 2014". Many of phishing techniques are sophisticated, and it is very hard to internet users to defend against them. Damage caused by phishing ranges from minor to substantial financial loss. According to the statistics provided by APWG in their Phishing Activity Trends Report[1] "Overall phishing activity was up by 20 percent in 3rd Quarter of 2013 from the previous quarter ", and Cyveillance whitepaper 2008 reported phishing attacks

against more than 2,000 brands across 30 countries which costs these organizations from thousands to millions of dollars per attack.

The phishing techniques usually involve impersonating legitimate web sites to submit personal information directly to the phisher, or using malicious software that sends victim's data without his knowledge. In a typical phishing attacks, the victim receives fraudulent email asking him to visit a web site and confirm his information in a given time. The email provides a legitimate-looking URL which direct to a spoofed web site where victims are going to enter their information.

Several of phishing solutions exist like blacklists which are databases of known phishing sites, whitelists, community ratings, analysis of the URLs and webpage content (images, and text), using machine learning techniques, and various heuristics to detect phishing attacks.

This paper makes the following contribution: We uses URL structure, four lexical features and page ranking to capture phishing attacks that depends on deceptive links. Every URL is evaluated corresponding to three heuristics (sub domain, primary domain, and path) and three lexical features extracted from the URL combined with page ranking received from ranking services. The proposed method requires no server changes and will prevent from phishing attacks based on fraudulent URLs. This solution uses resources like search engine suggestions, and third party services (Alexa, and Google Page Rank).

II. RELATED WORK

There are several methods that can be used to identify a web page as a phishing site, including Whitelists/Blacklists, URL and Heuristic-based, Similarity assessment techniques, and community ratings. In this section we will go through some of these solutions.

Whitelist/Blacklist-based is one of the common used approaches. It holds URL of verified phishing site. A whitelist contains URLs of legitimate sites while a blacklist contains phishing sites. It is effective to protect against phishing attacks and generates close-to-zero false positive rate but requires regular updating and is vulnerable to zero-day attacks. Many anti-phishing technologies rely on this approach. For example, Internet Explorer has built-in blacklist-based anti-phishing solution provided by Microsoft servers. Also Google's Safe Browsing extension which uses Google global blacklist and whitelist.

Content-based solutions which verify web pages by examining their contents (e.g. HTML, links, images, and text) against some previously defined characteristics. CANTINA [2] is an example on this approach which uses five words taken from the website to be classified as a signature using Term Frequency-Inverse Document Frequency (TF-IDF), then submits them to Google. If the site's URL is on top results it is legitimate, otherwise it is not. In CANTINA+ [3] which is an enhanced version from CANTINA, new features added and evaluated on a larger corpus to achieve better results. The new approach extended some of previous features combine them with ten more features. And this time the model built using state-of-the-art machine learning algorithms instead of a simple linear classifier. However, both of them have a drawback of time consumption caused by querying search engines.

A third approach developed to improve authentication between the user and the server. Authentication means that before user enters login information he needs to authenticate himself to that page. Also, it means that particular page authenticates to the user that it's the real page (called two way or mutual authentication). Some of anti-phishing techniques provide mutual authentication to prevent phishing attacks. This addresses the problem of user's inability to authenticate the website he is communicating with. The typical method used in login helps to authenticate user to server side but not the opposite which leaves a chance to attackers to exploit this failure. The success of mutual authentication techniques depends on the way used to authenticate both the client and the server. Some of existing solutions are image-based like the one provided by Confident Technologies company [4] which based on providing a number of categories instead of a specific pictures and let the user choose from them in registration process. At login, server will generate a grid of pictures and asks the user to choose the pictures matches the categories and order chosen in registration level. As soon as the server failed to provide the right grid of images or the user failed to choose the correct images it considered as security warning. Unfortunately, this maximize user's responsibilities by relying on user to memorize more than one category in specific order besides memorizing a password. Also, it requires changes on server side and login mechanisms. Other solutions uses Image-based user authentication to replace traditional methods (e.g. passwords, and security questions) this may provide stronger authentication but does not solve the server side authentication problem.

PwdHash [5] proposed a solution to strengthen web password authentication. It implements password hashing with domain name as a salt and keyed by the password itself. Server received password after hashing which makes it not useful if received by phishing website. As many of other solution this approach require user to remember using it every time he is about to enter a password.

Dhamija et al.[6] provide an authentication scheme where password is entered into a trusted window and user recognizes one image to perform visual matching to authenticate the received content. Images are generated by the server and they are unique for each transaction. The drawback of this solution is the large amount of changes required on server side.

BogusBiter[7] solves the problem from another point of view. It focuses on the stage after phishing attack occurred and user submitted his information to the wrong recipient. It automatically generates and sends a large number of fake credentials to phishing site to hide the real one. Unfortunately, BogusBiter can't work alone it needs help to be turned on from web browser or a third-party toolbar to detect phishing sites.

Aravind et al.[8] propose an anti-phishing framework which uses visual cryptography for authentication. Image is decomposed into two shares one stored with user and the other one is on website's database. An image captcha is created from those two shares in login time. The proposed method success to authenticate both user and website.

Web Wallet[9] is a sidebar login box which displayed when a user requested to login through a trusted path. It is responsible for preventing users from submitting their sensitive data directly to any website before checking that site. The developers of this sidebar used the negative visual feedback to solve the vulnerability of spoofing the sidebar and they provide cards to most of user's sensitive data not only user name and password.

TrueWallet[10] is another wallet-based approach which works as a proxy to manage user login and protect his password and credentials. It runs isolated from browser which adds an advantage to it compared to Web Wallet approach which means it is more secure and difficult to be attacked. TrueWallet uses the standard SSL-based authentication with some modification on server side. This approach has two disadvantages. First, it is vulnerable to DNS-spoofing attacks. Second, user need to be trained in order to rely only on this method to fill in any form.

One area of work relies on URL features to detect phishing webpage. Khonji et al.[11] propose a technique for detecting phishing websites by lexically analyzing suspect URLs depending on a novel heuristic phishing feature. This technique targets a subset of phishing attacks where the victim name is included in the URL. The approach achieved 63% and 83% true positive rate for loose and strict modes respectively. Whittaker et al.[12] present the design and evaluation of a large-scale machine learning based classifier. The proposed classifier evaluates the page according to its URL, content, and host information. The dataset used in training process consists of a noisy dataset of millions of samples. The evaluation concludes with more than 90% of phishing pages correctly identified.

An approach developed by Le et al.[13] to identify phishing target using only lexical features. Authors used an online method Adaptive Regularization of Weights in classifying URLs. Analysis showed that this methodology led to high classification accuracy comparable to full featured approaches. An approach that relies on 23 features derived from URL structure, lexical features, and from brand name of website is proposed in by Huang et al.[14]. These features model the SVM-based classifier used to inspect each requested URL. The evaluation done using three datasets containing more than 12,000 URLs and showed that the solution can obtain 99% accuracy.

Blum et al.[15] have proposed a method exploits URL's lexical features that are fed to the confidence-weighted algorithm, to indicate suspicious URL. This method uses a large lexical model trained using online approach which makes it capable of detecting zero hour threats. Zhang et al.[16] proposed different method based on repository to extract features and a statistical machine learning algorithm avoiding the complexity of computation caused by URL-based method. This method succeeds to identify phishing sites with more than 93% accuracy.

Nguyen et al.[17] presented a heuristic-based algorithm uses the characteristics of the URL combined with a third party services (e.g. PageRank) giving the URL a major role in phishing detection. Another classifier produced as a toolbar (PhishShark) proposed which is heuristic-based-only combining URL and HTML features led to promising results.

Finally, we will conclude with some of existing toolbars[18] built to prevent phishing attacks. Netcraft is a Mozilla browser plug-in that displays host location and risk rating of the accessed site. User can report sites to Netcraft to validate them then add them to its blacklist database if they are phished. TrustWatch is toolbar for Internet Explorer that checks the URL in the black listed database and displays its domain name. Searching blacklist is a time consuming process since they continuously growing and they are vulnerable to zero-day-attacks. Spoofguard is an anti-phishing Internet Explorer plug-in. It examines page characteristics such as images, links, and domain name against common features extracted from phishing site to decide whether this page is spoofed or not.

III. PROPOSED APPROACH

Our system is inspired by solutions proposed by Nguye et al. [18] and Xiaoqing et al. (GU Xiaoqing, 2013). It combines both approaches the heuristic-based approach and NB classifier. In Nguye et al. solution URL-related features and Page Ranks used to classify each website. Xiaoqing et al. approach depends onto two phases. The first one is an NB classifier which uses four lexical features to decide whether URL is phishing, suspicious, or legitimate. The second phase uses SVM classifier to parse the webpage against some features. The system will enter the second phase only if URL classified as suspicious in first phase. In our proposed system we combined the first approach with the first phase of second approach without entering into its second phase.

A. System Model

Our system model consists of seven main modules as illustrated in figure 1.

- Receiving URL Module

The system obtains the requested URL from the browser. The output of this module is page URL and it is a fundamental input in most of system modules.

- Features Extraction Module

This module extracts URL domain-related features. The URL separated into different components which are Primary Domain, Sub Domain, and Path. The pervious features will

play an eminent role for investigating the URL and predicting phishing pages in next modules.

- Ranking module

Beside URL's feature extraction in previous modules, also this module collects URL metadata. Specifically, URL Page Rank to be used as input into next module. Google Page Rank, and Alexa Rank are used for this purpose.

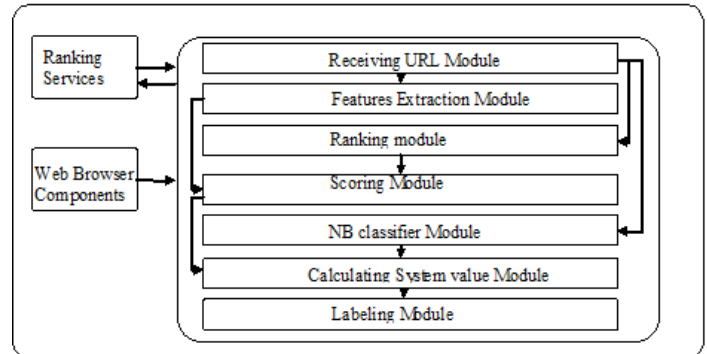


Fig. 1. System Modules

- Scoring Module

In this module heuristics derived from modules B and C are used as input and their values calculated as output. As a result, the site is considered as phishing if all calculated values are negative, and is considered as legitimate if they are all positives.

- NB classifier Module

This module is responsible for classifying a URL with a classification model developed in training process. The features used by the classification system are checking whether the URL contains an IP address because this method used by phishers to hide the owner of the site. Another feature is to examine the presence of a large number of dots separating hostname. Phishers tend to use more dots in their URLs to impersonate a legitimate look of URL because there is no restrictions on the number of dots can be used in sub domains. Checking URL against special symbols such as '@' or '-' is another feature because many of phishing URLs modified using these symbols which makes it possible to write URLs that appear legitimate but actually lead to different pages. URLs corresponding to legal websites usually do not have a large number of slashes [19]. As a result, URL that contains a large number of slashes is considered to be a phishing. The classifier entered into two phases training, and testing phases. Training phase used to build the classifier by calculating the probabilities that the given webpage belongs to a one of two classes (phishing, and legitimate). The testing phase is used to examine the ability of classifier to label real web pages with a correct class.

- Calculating System Value Module

In this step each heuristic is given a weight obtained by a classifier. After that, system values are calculated using this equation:

$$VS = \sum_{i=1 \text{ to } 6} (\text{heuristic}_i \text{ value}) * (\text{heuristic}_i \text{ weight})$$

- Labeling Module

This module deals with system value and compares it to threshold to give system output which is the URL final label. As a result, user may proceed safely, or warned about the website.

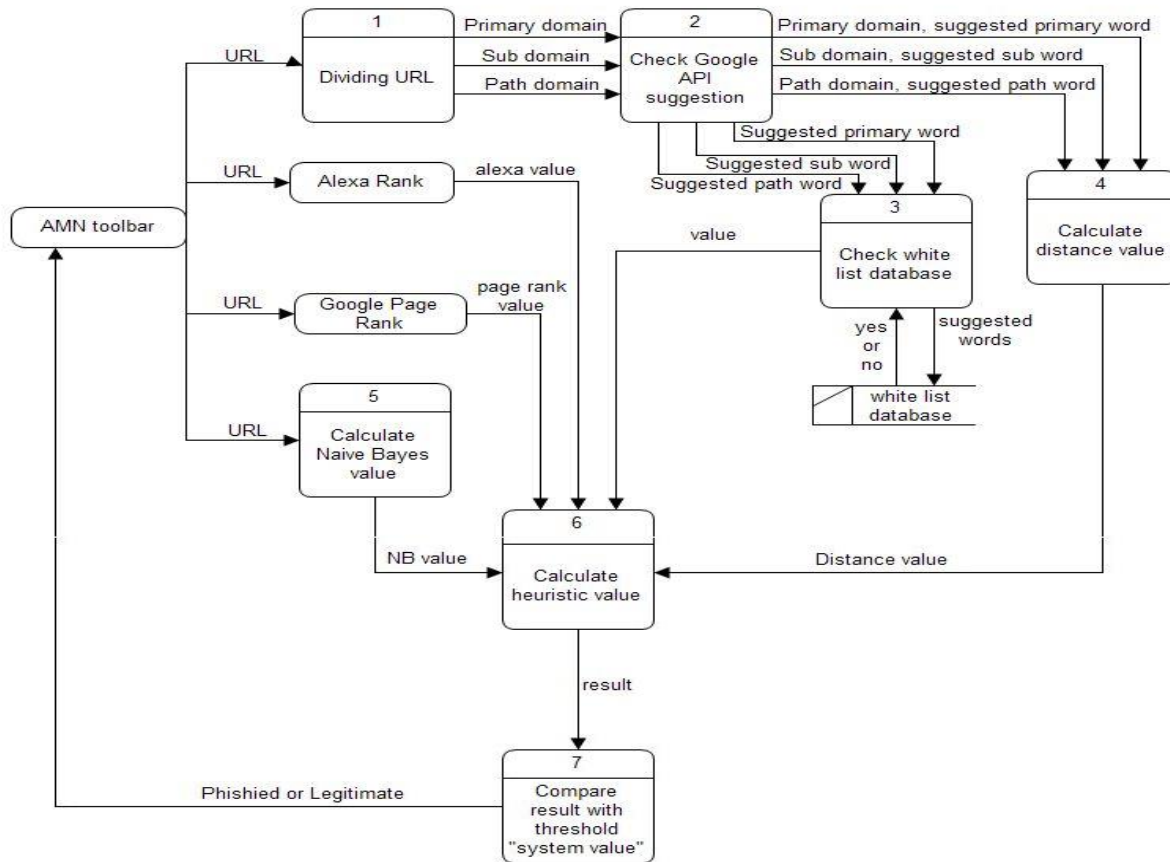


Fig. 2. Data Flow Diagram

B. Structured Design

It shows data exchanged between system components. As we can notice, from the Figure-2 URL is the main part of data. Most of major components and processes need URL value as input to produce their results. Also, URL in most cases need to be decomposed into three parts (sub domain, primary domain, and path) which is the responsibility of "Dividing URL" process. Process two receives URL parts and returns suggestions of each part separately. Suggestion values used by two processes three and four to check them in a list of popular phishing targets as in process three then return a value of yes or no. Process four produces edit distance value between each suggested word and its corresponding URL part. Process six is the major part of the system since all of the data produced in other system processes will be used here to calculate final system result. Process seven is the last process in system which communicates only with one process to receive result value and compare it with a predefined threshold to make system decision.

C. Proposed Algorithm

The pseudo code of our proposed algorithms to detect phishing websites are described below.

Algorithm Primary Domain Value

```

Input Primary Domain
Output Primary Domain Value
if (Primary Domain=null)
    Value= -0.5
else
    S=Suggestion(Primary Domain)
if (S=null)
    Value= -0.25
else
    T=Whitelist(S)
    if (T)
        Ed=Levenshtein(S, Primary Domain)
        if (Ed=0)
            Value= 1
        else-if (0< Ed<3)
            Value= -0.5
        else-if ( Ed>=3)
            Value= 0.25
    else
        Ed=Levenshtein(S, Sub Domain)
        if (Ed=0)
            Value= 0.5
        else-if (0< Ed<3)
    
```

```
Value= -0.25
else-if ( Ed>=3)
Value= 0.25
```

Algorithm Sub Domain Value

```
Input Sub Domain
Output Sub Domain Value

if (Sub Domain=null)
Value= 0
else
S=Suggestion(Sub Domain)
if (S=null)
Value= 1
else
T=Whitelist(S)
if (T)
Value= -1
else
Ed=Levenshtein(S, Sub Domain)
if (Ed=0)
Value= -0.5
else-if (0< Ed<3)
Value= -0.25
else-if ( Ed>=3)
Value= 0.5
```

Algorithm Path Domain Value

```
Input Path Domain
Output Path Domain Value

if (Path Domain=null)
Value= 0
else
S=Suggestion(Path Domain)
if (S=null)
Value= 1
else
T=Whitelist(S)
if (T)
Value= -1
else
Ed=Levenshtein(S, Path Domain)
if (Ed=0)
Value= -0.5
else-if (0< Ed<3)
Value= -0.25
else-if ( Ed>=3)
Value= 0.5
```

Algorithm Page Rank Value

```
Input Page Rank
Output Page Rank Value

if (Page Rank<=0) then Value= -1
if (1<=Page Rank<=2) then Value= -0.5
if (3<=Page Rank<=4) then Value= -0.25
if (5<=Page Rank<=6) then Value= 0.25
if (7<=Page Rank<=8) then Value= 0.5
if (9<=Page Rank<=10) then Value= 1
```

Algorithm Alexa Rank Value

```
Input Alexa Rank
Output Alexa Rank Value

if (Alexa Rank<300,000) then Value= 1
if (300,000<= Alexa Rank<=500,000) then Value= 0.5
if (500,000<= Alexa Rank<=1000,000) then Value= 0.25
if (1000,000<= Alexa Rank<=2000,000) then Value= -0.25
if (2000,000<= Alexa Rank<=3000,000) then Value= -0.5
if (Alexa Rank>=3000,000) then Value= -1
```

Algorithm NB Classifier

```
Input URL
Output Class Value
host=Host(URL)
path=Path(URL)
features[x1, x2, x3, x4]=Extract(host,path) //Each feature xi
takes value 0, or 1
 $P(C_p | \text{features}) = P(C_p) * \prod_{i=1 \text{ to } 4} P(x_i | C_p)$  //  $C_p$  is class phishing
 $P(C_l | \text{features}) = P(C_l) * \prod_{i=1 \text{ to } 4} P(x_i | C_l)$  //  $C_l$  is class legitimate

if ( $P(C_l | \text{features}) / P(C_p | \text{features}) > \alpha$ ) Class=1 //legitimate
else-if ( $P(C_p | \text{features}) / P(C_l | \text{features}) > \alpha$ ) Class=-1
//phishing
else-if ( $(1/\alpha) < P(C_l | \text{features}) / P(C_p | \text{features}) < \alpha$ ) Class=0
//suspicious
```

Algorithm Calculate System Value

```
Input Heuristic Values and Weights
Output Website Final Class

 $VS = \sum \text{heuristic value} * \text{heuristic weight}$ 

if ( VS < Threshold ) Value= 0 // Phishing class
else Value= 1 // Legitimate class
```

• ALEXA RANK

It is a service from Amazon Company since 1996, which gives a value for each page through 3 months in the Web. Increasing of this value is a good indicator. The value depends on 2 important things. First, the number of unique users entered to this site. Second, how many URL linked to this site, increasing of URL's lead to this site will increase its value (Alexa API).

This service serve project to detect phishing sites, because phishing sites has a few number of visitors and linking URLs compared to popular websites. Also, phishing sites usually have a short life cycle which helps to differentiate between legitimate and phishing sites.

• PAGE RANK

It is a service from Google Company. When Google needed to improve searching on web by giving best results to searchers, they thought about giving a value for each page (Karch). High values depend on how many URL linked to the site. Also the value depends on the domain age, the older

domain get higher value (Strickland, 2006). So the proposed approach, used this value to be one of factors that affect the decision about whether the site is phishing or not.

- SUGGESTIONS

When user enters to "Google.com" and type a word, there is a drop down list to suggest many words related to user's typed word. The suggestions depend on word popularity in searching. And when you enter a word spelled wrong, there is a famous sentence says "Did you mean?" depends on common spellings (Autocomplete).

Since phishers try to make phishing URL similar to popular sites by adding some letters, removing others, or even substituting them with different letters to trick users that it is their targeted site. So, we used Google suggestions by taking the suspect URL and getting the relative spelling word, then compare those two words using levenshtein distance algorithm.

- LEVENSHTTEIN ALGORITHM

It is an algorithm that compares two strings and returns the number of operations (insert, delete, and substitute) known as "distance" to let these words sound the same.

Since our paper use this algorithm to compare between suspect word and Google suggested word and return the number of operations to let those two words be equivalent. If the distance is 0, this means the two strings are the same. But if the distance is between 1 and 2 that means a probability of some phisher is trying to make those two words visually similar.

- WHITE LIST

Usually in phishing world, the white list is group of legitimate sites saved in database. But in our proposed algorithm white list means a list saving primary domains of sites targeted by phishers. This white list is extracted from a database of verified phishing URLs downloaded from PhishTank website. We need to check if the URL domains (primary domain, sub domain and path domain) not in the white list to ensure that there is no phisher exploits the name of a famous legitimate site to trick users.

D. Functions Implementation

Function Alexa Rank

Prototype: Function Alexa (url)

Input: URL.

Output: URL's value.

Description: This JavaScript function takes the URL as a parameter, and connects to the server using AJAX to send URL to PHP file which requests Alexa rank API for the URL. Then receive URL's global rank from the server. After that it assigns URL a value based on its rank. Whenever rank is higher, assigned value becomes bigger.

Function Sub Domain

Prototype: Function S_heuristics (SubDomain)

Input: Sub domain.

Output: Sub domain value.

Description: This JavaScript function takes URL's sub domain and passes it to three functions to compute sub domain value, those functions are:

1) *Google's search suggestions:* return the Google suggested word for the sub domain.

2) *White list:* check whether the suggested word is a primary domain of another targeted site. If it is in the white list, sub domain will assigned a low value.

3) *Levenshtein algorithm:* if the sub domain is not in the white list, this function will check the distance between sub domain and the suggested word to check if the sub domain attempts to be closed to another domain. Whenever the distance is lower, the value becomes higher.

Function White List

Prototype: Function whitelist(a)

Input: Google suggestion word.

Output: True or False.

Description: This JavaScript function connects to "PhishTank" database. Each phish site in database has phish id, phish URL, target and other columns. Important columns are:

* Phish URL: The URL of the phishing site.

* Phish id: Id for each phish URL.

* Target: The primary domain of legitimates site which are the phish site attempts to simulate. JavaScript function passes the Google suggested word to PHP file using Ajax to create connection to the database to check whether the suggested word matches any target. If it exists that means this site attempt to disrupt the user to think it is the primary domain of another legitimate site. The returned value in matching case is true. If the result is true, white list will take low value.

Function NB Classifier

Prototype: Function NB_classifier(host, path)

Input: URL host and path.

Output: Class value.

Description: This function implemented based on Naïve Bayes classification approach. It is a learned classifier trained over a data set of 12,967 phishing URL downloaded from PhishTank and 150 legitimate URL collected manually with help of Alexa top 500 URLs. Features used for classification illustrated in table 1.

TABLE I. FEATURES USED BY NB

| Heuristic | Phishing URL |
|----------------|-------------------------|
| Suspicious URL | URL contains @ or - |
| IP Address | URL contains IP address |
| Dots in URL | >=5 dots in URL |
| Slash in URL | >=5 slashes in URL |

These features extracted from each URL. Finally, classifier will return one of three values (0, 1, or -1) suspicious, legitimate, or phishing, respectively.

Function Calculation

Prototype: Function calculation()

Input: Nothing.

Output: System Value.

Description: This function is the main part of the program and the last step of calculations. It applies equation $vs = \sum (\text{heuristics value}) * (\text{heuristics weight})$, where heuristic values are taken from global variables used to store results of previous functions. And heuristic weight calculated by experiments applied on phishing URLs. The result of these function vs returned to calling function to be compared with threshold before presenting the last decision of the program.

IV. PERFORMANCE EVALUATION

Our proposed architecture for anti phishing toolbar uses an extended approach from Nguye et al. [17] by combining their approach with NB classifier proposed in X. Gu, et al. [19]. Algorithms illustrated bellow are based on experimental results of 9,661 phishing URL downloaded from PhishTank as Nguye et al. mentioned. Naïve Bayes classifier algorithm used for classification trained over 12,967 phishing URL from PhishTank and 253 legitimate URL collected manually.

Evaluation phase is done in three phases as shown in Table 9. The dataset used for testing is collected using two methods from PhishTank, and manually. URLs in data sets evaluated manually by installing the toolbar and testing each URL individually. Metrics used to calculate toolbar accuracy are True Positive (classifying legitimate URL correctly), False Positive (assigning phishing label to legitimate URL), True Negative (predicting phishing site correctly), and False Negative (assigning legitimate label to phishing URL).

First phase, we started by evaluating Naïve Bayes (NB) approach alone. NB classifier trained over 13117 URLs divided into 12967 phishing URLs, and 150 legitimate URLs. After that, NB tested using a test set of 13220 URLs (12967 phishing, 253 legitimate). We experimented NB using different values of α as illustrated in Table 3. The best value of α which maximizes TN and minimizes FP is 5.8.

In the second phase we evaluate system without Naïve Bayes addition. The test set consists of 162 URLs to be tested. Experiment are done using threshold of value 0. Toolbar detected 77 phishing URLs correctly out of 89, and 63 legitimate URL out of 71. The experiment results with 86.5% True Negative, 88.7% True positive, 11.2% false positive, 13.4% false negative.

Third phase, we combined both of previous approaches. The test set consists 156 phishing URLs (selected from 12,967 URLs) downloaded from PhishTank, and 90 legitimate URLs collected manually. We conclude with 246 URLs developed for testing. We experiment this approach using two different values of threshold 0, and 0.5. Threshold with 0.5 results with less False Negative, so we select it as threshold value. Although it returns a high False Positive it gives a good results of True Negative. False Positive can be reduced using "add to trusted list" feature. The toolbar detected 147 phishing URL correctly out of 156. The experiment results with 94% True Negative.

Accuracies of each approach calculated using this equation: Accuracy ratio= (TP+TN)/(TP+TN+FP+FN). Phase one has 34% accuracy ratio, phase two has 87.5% , and phase three has 48% accuracy ratio.

Finally, after these experiments we concluded by choosing threshold of value 0, and remove the Naive Bayes part as it does not add any improvement on system accuracy and increases false positive rate.

TABLE II. EVALUATION PHASES

| Phase 1: | Naïve Bayes | Trained on | 13117 (12967 P + 150 L) URL | | | | |
|----------|------------------------|------------|---------------------------------------|-------|-------|-------|-------|
| | | | TP | TN | FP | FN | |
| | | Tested on | 13220 (12967 P + 253 L) URL | 248 | 4257 | 5 | 8710 |
| | | | | 98% | 33% | 2% | 67% |
| Phase 2: | URL-based Approach | Tested on | 162 (89 P + 71 L) URL Threshold=0 | 63 | 77 | 8 | 12 |
| | | | | 88.7% | 86.5% | 11.2% | 13.4% |
| Phase 3: | Combination of 1 and 2 | Tested on | 246 (156 P+90 L) URL Threshold=0.5 | 10 | 147 | 80 | 9 |
| | | | | 11% | 94% | 88% | 6% |

TABLE III. α Values

| α | A | TN | FP | A | TN | FP |
|----------|-----|------|----|-----|------|----|
| | 1.1 | 6058 | 18 | 3.5 | 6058 | 18 |
| | 1.3 | 6058 | 18 | 3.9 | 6058 | 18 |
| | 1.5 | 6058 | 18 | 4.4 | 6058 | 18 |
| | 1.7 | 6058 | 18 | 5.6 | 6058 | 18 |
| | 2.0 | 6058 | 18 | 5.7 | 6058 | 18 |
| | 2.4 | 6058 | 18 | 5.8 | 4257 | 5 |
| | 2.6 | 6058 | 18 | 5.9 | 4257 | 5 |
| | 2.9 | 6058 | 18 | 6 | 4257 | 5 |
| | 3.2 | 6058 | 18 | 6.3 | 4257 | 5 |

V. CONCLUSION

This paper presents architecture for developing an anti-phishing toolbar integrated with Firefox browser to detect phished URLs. Our proposed anti-phishing toolbar will verify user's inputted URL, if the result is phished then it warns the user through changing indicator color and gives the user the choice to unblock the website by adding it to a trusted list. In case of phishing site verified user can know the reason by viewing a report. Our approach categorizes the URL based on its features including four lexical features and three other features (sub domain, primary domain, and path) with help of Naïve Bayes classifier. Our proposed approach can minimize the false positive by giving the user a feature of adding URLs after verified to a trusted list. Experimental results show that our approach can achieve 48% accuracy ratio using a test set of 246 URL, and 87.5% accuracy ratio by excluding NB addition tested over 162 URL.

REFERENCES

- [1] Anti-Phishing Working Group. "Phishing Activity Trends Report: Third Quarter 2013 Report," April 2014. Available : http://docs.apwg.org/reports/apwg_trends_report_q3_2013.pdf.
- [2] Y. Zhang, J.I. Hong, L. F. Cranor, " Cantina: A Content-Based Approach to Detecting Phishing Web Sites," In Proceedings of the 16th International Conference on World Wide Web (WWW '07). ACM, NY, USA, 639-648, 2007.
- [3] G. Xiang, J. Hong, C. P. Rose, L. Cranor, "CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites," in ACM Transactions on Information and System Security (TISSEC), Volume 14 Issue 2, September 2011.
- [4] Confident Technologies, "Dynamic, Mutual Authentication Technology for Anti-Phishing" Available: <http://confidenttechnologies.com/products/anti-phishing>.
- [5] B. Ross, C. Jackson, N. Miyake, D. Boneh, J. C. Mitchell, " Stronger password authentication using browser extensions," In Proceedings of the 14th conference on USENIX Security Symposium - Vol. 14. USENIX Association, Berkeley, CA, USA, 2005.
- [6] R. Dhamija and J. D. Tygar, " The battle against phishing: Dynamic Security Skins", In Proceedings of the 2005 symposium on Usable privacy and security (SOUPS '05). ACM, NY, USA, 77-88. 2005.
- [7] C. Yue and H. Wang , "BogusBiter: A transparent protection against phishing attacks," in ACM Transactions on Internet Technology (TOIT), Volume 10, Issue 2, May 2010 .
- [8] K.A.Aravind, R.M. Venkata Krishnan, "Anti-Phishing Framework for Banking Based on Visual Cryptography," International Journal of Computer Science and Mobile Applications, Vol. 2 Issue. 1, January, 2014.
- [9] M. Wu, Robert C. Miller, G. Little., "Web wallet: preventing phishing attacks by revealing user intentions," In Proceedings of the second symposium on Usable privacy and security (SOUPS '06). ACM, New York, USA, 2006.
- [10] S. Gajek, H. Löhr , A. R. Sadeghi, M. Winandy. , "TruWallet: trustworthy and migratable wallet-based web authentication," In Proceedings of the ACM workshop on Scalable trusted computing (STC '09). ACM, NY, USA, 19-28, 2009.
- [11] M. Khonji, A. Jones,Y. Iraqi, "A Novel Phishing Classification Based on URL Features," in IEEE GCC Conference and Exhibition (GCC), Dubai, 2011.
- [12] C. Whittaker, B. Ryner, M. Nazif , "Large-Scale Automatic Classification of Phishing Pages," in Network and Distributed System Security Symposium (NDSS), 2010.
- [13] A. Le, A. Markopoulou, M. Faloutsos, "PhishDef: URL Names Say It All," INFOCOM 2011,Shanghai, China, 191-195, 2011.
- [14] H. Huang, L. Qian and Y. Wang , "A SVM-based Technique to Detect Phishing URLs," Information Technology Journal, Volume 11 Issue 7, page 921-925, 2012.
- [15] Blum, B. Wardman, T. Solorio, G. Warner, "Lexical feature based phishing URL detection using online learning" In Proceedings of the 3rd ACM workshop on Artificial intelligence and security (AISec '10). ACM, NY, USA, 54-60 , 2010.
- [16] J. Zhang, Y. Wang, "A real-time automatic detection of phishing URLs," Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on, vol., no., pp.1212,1216, 29-31 Dec. 2012.
- [17] L. A. T. Nguyen, B.L. To, H. K. Nguyen, M. H. Nguyen , "Detecting Phishing Web sites: A Heuristic URL-Based Approach," in International

- Conference on Advanced Technologies for Communications (ATC'13), Oct. 2013.
- [18] L. F. Cranor, S. Egelman, J. Hong, Y. Zhang, "Phishing phish: Evaluating anti-phishing tools," in 14th Annual Network & Distributed System Security Symposium (NDSS 2007), 2007.
- [19] X. Gu, H. Wang, T. Ni , "An Efficient Approach to Detecting Phishing Web," Journal of Computational Information Systems 9, pp. 5553-5560, 2013.
- [20] "Alexa API," [Online]. Available: <http://data.alexacom/data?cli=10&dat=snbamz&url>.
- [21] M. Karch, "What Is PageRank and How Do I Use It?," [Online]. Available: <http://google.about.com/od/searchengineoptimization/a/pagerankexplain.htm>.
- [22] J. Strickland, "How Google Works," 20 December 2006 . [Online]. Available: <http://computer.howstuffworks.com/internet/basics/google1.htm>.
- [23] "Autocomplete," [Online]. Available: <https://support.google.com/websearch/answer/106230?hl=en>.

A Study of Scala Repositories on Github

Ron Coleman

Computer Science Department
Marist College
Poughkeepsie, New York, United States

Matthew A. Johnson

Computer Science Department
Marist College
Poughkeepsie, New York, United States

Abstract—Functional programming appears to be enjoying a renaissance of interest for developing practical, “real-world” applications. Proponents have long maintained that the functional style is a better way to modularize programs and reduce complexity. What is new in this paper is we test this claim by studying the complexity of open source codes written in Scala, a modern language that unifies functional and object programming. We downloaded from GitHub, Inc., a portfolio of mostly “trending” Scala repositories that included the Scala compiler and standard library, much of them written in Scala; the Twitter, Inc., server and its support libraries; and many other repositories, several of them production-oriented and commercially inspired. In total we investigated approximately 22,000 source files with 2 millions lines of code and 223,000 methods written by hundreds of programmers. To analyze these sources, we developed a novel compiler kit that measures lines of code and adaptively learns to estimate the cyclomatic complexity of functional-object codes. The data show, first, lines of code and cyclomatic complexity are positively correlated as we expected but only weakly which we did not expect with Kendall’s $\tau=0.258-0.274$. Second, 75% of the Scala methods are straight-line, that is, they have the lowest possible cyclomatic complexity. Third, nearly 70% of methods have three or fewer lines. Fourth, the distributions of lines of code and cyclomatic complexity are both non-Gaussian ($P<0.01$), which is as surprising as it is interesting. These data may offer new insights into software complexity and the large-scale structure of applications including but not necessarily limited to Scala.

Keywords—Functional programming; Scala; GitHub.com

I. INTRODUCTION

Functional programming appears to be enjoying a renaissance of interest for writing practical applications. The turn toward functional programming is evident in recent introductions of new functional languages, revival of old ones, incorporation of functional semantics in non-functional languages, publications of trade texts focused on functional programming, proliferation of open source communities and tools dedicated to functional programming, and adoption of functional approaches by some firms in industry. While reasons for the newfound enthusiasm are likely varied, some proponents have argued that elaboration of the lambda calculus is well suited to writing modular programs that reduce complexity.

What is new in this paper is we tested this latter claim and developed an experimental compiler kit to analyze the modularity and complexity of Scala, a modern language that unifies functional and object styles [1,2]. (While the focus is on Scala, the concept we present is more general and we posit adaptable to the functional style, whether in a pure functional

language or a language like Java that recently incorporated lambda expressions beginning with version 8.) We then downloaded from GitHub.com¹ a portfolio of mostly “trending” Scala repositories that contain millions of lines of source in tens of thousands of files with hundreds of thousands of methods written by hundreds of programmers. A robust analysis of this data indicates that lines of code (LOC) and cyclomatic complexity (M) [3] are positively correlated, as we expected, but only weakly which we did not expect. In other words, LOC and M are clearly related though not necessarily interchangeable as suggested elsewhere in the literature for programs written with imperative languages. While we do not yet know if this new finding is unique to Scala, robust variability statistics indicate M is a more reliable estimate of complexity compared to LOC, confirming the distinction of the two metrics, at least for the Scala repositories on GitHub. The data furthermore shows an interesting non-Gaussian preponderance of short, straight-line methods, which also surprised us. That is, we assumed as a null hypothesis that LOC and M would be normally distributed about a mean value which they aren’t. These new findings may offer insights into software complexity and the large-scale structure of programs including, but not necessarily limited to, Scala.

II. BACKGROUND

Functional programming for much of its history has thrived largely in academic obscurity [4,5]. That may be changing. A renaissance of interest in “real-world” applications of functional programming, languages, and styles has emerged in recent years [6,7,8,9,10,11,12]. Some investigators have observed that the renewed enthusiasm for functional programming is partly a response to the “free lunch is over” dilemma posed by the advent of commodity multicore systems [13,14]. Others like Twitter, Inc., have switched to functional programming, and Scala in particular, for the advantages Scala purports to offer for scalability.² Yet functional programming proponents have long maintained that mathematical expressiveness of the functional style lends itself to modularization and reducing program complexity [15]. That there have been no empirical studies to support these latter suppositions has not stopped language designers, developers, and authors from arguing for more functional programming.

We don’t fault functional programming enthusiasts. There isn’t even a consensus regarding what software complexity *is*, a conundrum in our view reminiscent of asking what beauty *is*,

¹See GitHub, Inc., <https://github.com/trending?l=scala>, accessed: 6 June 2014

²See C. Metz, “Twitter jilts Ruby for Scala,” The Register, http://www.theregister.co.uk/2009/04/01/twitter_on_scala/, accessed 4 Jun 2014

which Immanuel Kant tackled more than two centuries ago [16]. Perhaps the relationship between software complexity and aesthetics and matters of taste is more than a philosophical one for if software complexity were “in the eyes of the beholder” it might account for the 100-plus different metrics, computational and cognitive, that propose to blindly quantify what is desirable and undesirable in code, the irony notwithstanding [17,18]. Our point is only to suggest that rather than inventing yet another metric for functional programs, we believed it more productive as an experiment to start with existing metrics, off-the-shelf, so to speak; refactor them only if needed; and see what the source code is telling us.

As a candidate, M has its downsides, being imperfect and dated [19,20]. Furthermore there is no research on how to apply M to functional programs, which differ in some fundamental ways from imperative programs for which M had been originally developed. Still M remains the most widely known and often applied metric, standing singularly for its diverse implementations³ and published risk assessments by the Software Engineering Institute [21].

LOC, as a simple measure of complexity, is similarly dated and inadequate [22]. Some modern languages furthermore present semantic challenges for measuring LOC because of nested definitions and structures. Nevertheless the enduring importance LOC, despite their limitations, are evident in modern source editors, IDEs, operating systems, etc. which would be incomplete from a programming point of view without line counting facilities.

Hatton observed for FORTRAN and C that M and LOC were statistically correlated, declaring M “effectively useless” [23]. Perhaps Hatton made this claim because LOC was so obvious and simple that there had to be a better approach, although for our purposes we show this view of LOC is naïve at best. We don’t disagree with Hatton in principle; we would simply state the matter differently. Namely, we expect only as a working hypothesis that any other measure of software complexity is positively correlated with LOC since this view comports with commonsense and anecdotal experience.

III. WHY SCALA?

We were motivated to study Scala for a few reasons.

1) *Scala blends functional and object-oriented styles, which stood out for us as representative of the forward-looking, modern turn toward practical functional programming.*

2) *Scala is a Java Virtual Machine (JVM) language. This is complementary to the first item above and it means Scala*

runs virtually everywhere (e.g., desktops, browsers, cellphones, tablets, and GPUs [24], and furthermore interoperates with a large installed base of legacy Java codes. Thus, a study of Scala may be of interest to a broader audience of programmers and researchers.

3) *Scala repositories are readily accessible as open source. Some of these repositories, as we show, are large, sophisticated, and deployed in a commercial / production capacity.*

4) *The Scala open source community provided us with the requisite tools to develop our own tools. We are referring mainly to the Scala plugin for Eclipse (see below) that was created and over the years improved by the Scala community.*

5) *We are Scala programmers. We have used Scala for teaching and research purposes and we were thus curious to know how our anecdotal experience compared with empirical data.*

IV. SCLASTIC

The main problems, conceptual and programming, were how to apply LOC and M to Scala. In summary, the issue for LOC is how to interpret inner definitions. The issue for M is handling standard library and user-defined high-order predicate functions.

Thus, we sought to implement an experimental compiler kit capable of solving these problems for a large sample of Scala source codes. We call this kit, *Sclastic* since it operates in an “elastic” manner, that is, it learns to dynamically estimate M by discovering the signatures of high-order functions that take predicate (i.e., Boolean-returning) function objects as parameters and storing this information in a database that *Sclastic* consults during a separate pass.

Sclastic is itself mostly written in Scala and the source is hosted on GitHub. ⁴ *Sclastic* is *not* in the portfolio of repositories we analyze.

At a high level, the main body of *Sclastic* has three phases, each comprising one or more passes. The first phase, the decommenter, removes comments and empty lines from the input file, which it stores as an in-memory stream of string objects. The second phase, the parser, filters the string stream and identifies lexical objects, methods, scopes, and *decision points*, which we describe below. The final phase, the method compiler, analyzes the parse stream and calculates lines counts and estimates M for each method.

This three-phase process outputs a list of Scala objects that other parts of the *Sclastic* kit analyze for statistical and report generation purposes.

A. Definitions

For the purposes of this paper, we have the following definitions.

- **Line.** A Scala line is a sequence of zero or more characters terminated by a newline character or the end

³ For only a smattering of languages see F. Kline, “Cyclomatic Complexity Viewer,” <http://www.codeproject.com/Articles/10705/Cyclomatic-Complexity-Viewer>, accessed 12 Aug 2013; G. Wilson, “Cyclomatic complexity for Python code,” <http://thegarywilson.com/blog/2006/cyclomatic-complexity-for-python-code/>, 9 Jul 2006, accessed 6 Jun 2014; G. Wilson, “Cyclomatic complexity for Python code,” <http://thegarywilson.com/blog/2006/cyclomatic-complexity-for-python-code/>, 9 Jul 2006, accessed 6 Jun 2014; SonarSource, <http://www.sonarqube.org/>, accessed 6 Jun 2014; and Cyvis, “Software Complexity Visualiser,” http://cyvis.sourceforge.net/cyclomatic_complexity.html, accessed 6 Jun 2014

⁴See “*Sclastic*,” <http://github.com/roncoleman125/Sclastic>, accessed 6 June 2014

of file. In the degenerate case, an empty line has only a newline character or is the end of file. A line may contain one or more comments. A de-commented line is a line with all comments removed.

- **Method.** A method is a Scala class member function. The function may return an object or it may be void returning in which case it may also be called a procedure. The method may have zero or more formal parameters. Every method is composed of at least one non-empty line.

B. Line counts

Per the definition above, counting lines in Scala source, in the simplest cases, is simple. Consider the snippet below.

```
1: class A {  
2:   def evens(input: List[Int]):  
      List[Int] = {  
3:     input.filter(p => p % 2 == 0)  
4:   }  
5: }
```

Snippet 1

The `evens` method of class `A`, given a list of integers, returns a new list with only the even numbers from the input list. (Note: Scala ignores indents and most whitespace. We added line numbers only for readability; they are not part of Scala syntax.) In this case, the line count of `evens` which Sclastic reports, is the number of lines between and including the inner curly braces of `evens`, that is, three.

Curly braces, however, are often optional in Scala. Consider the snippet of class `B` below.

```
1: class B {  
2:   def evens(input: List[Int]) =  
      input.filter(p => p % 2 == 0)  
3: }
```

Snippet 2

This `evens` method is functionally equivalent to the one from class `A`. However, Sclastic compiles `evens` in this case giving a line count of one.

Counting lines is even subtler since Scala permits inner definitions of classes and methods. That is, it is possible to define classes within classes, methods within methods, and combinations thereof with arbitrary nesting depth. Consider the snippet below, which implements `evens` with the closure, `iseven`.

```
1: class C {  
2:   val TWO = 2  
3:   def evens(input: List[Int]):  
      List[Int] = {  
4:     def iseven(a: Int): Boolean =  
          a % TWO == 0  
5:     input.filter(p => iseven(p))  
6:   }  
7: }
```

Snippet 3

While the “nominal” line count of `evens` is four (i.e., lines 3 – 6), the “effective” line count is three (i.e., lines 3, 5 and 6). The nominal and effective line count of `iseven` is one (i.e., line 4).

Note furthermore that class `C` has an implied constructor which initializes the member, `TWO`, whenever an instance of `C` comes into existence. (In Scala, a `val` type is a read-only “value” or constant, `final` in Java.)

Sclastic interprets constructors as initializer methods. Thus, while the constructor’s nominal line count is seven (i.e., lines 1 – 7), its effective line count is three (i.e., lines 1, 2, and 7).

We define the nominal line count to be the number of lines of a lexical scope including inner definitions. The effective line count is the number of lines of a lexical scope not including inner definitions. For purposes of this paper, we use only the effective line count.

C. Hard signatures

The cyclomatic complexity given by McCabe [5] is

$$M = E - N + 2P \quad (1)$$

where E and N are the number of edges and nodes, respectively, in the program flow graph and P is the number of exit points for a given method.

A simplification is to use predicate counting [5,33]. It counts *decision points*, i.e., statements that contain Boolean expressions where an alternate path though the code might be selected. If π is a function, which returns the number of decision points within a method, then the cyclomatic complexity can also be calculated as follows:

$$M = \pi + 1 \quad (2)$$

M McCabe [5] shows that Equations 1 and 2 give the same result for a method where $P=1$. In both cases, the smallest value, $M=1$, means the method consists only of a “basic block” or “straight-line” code.

Using Equation 2 makes calculating M straight forward for languages like Java. In this case, the decision point signatures are the selection and looping statements: *if*, *switch-case*, *for*, *while*, and *do-while*. Since these statements may also contain Boolean expressions connected by logical-and and/or logical-or operators, respectively, `&&` and `/ /` are also decision points in the context of selection and looping statements. For instance, we count an *if* statement as one decision point while we count an *if* statement with an embedded `&&` or `||` as two decisions points.

There is yet another simplification which we assume. Namely, if the Boolean expression in the selection or loop is constant true (i.e., there is no decision), it is still counted as a decision point, even though the Boolean expression will never be false.

As the table below suggests, there is an incomplete correspondence between Java and Scala decision point signatures.

TABLE I. JAVA AND SCALA DECISION POINT SIGNATURES

| Java | Scala |
|-------------|------------------------|
| if | if |
| for | for (possibly) |
| while | while |
| do-while | N/A |
| switch-case | match-case |
| &&, | &&, |
| N/A | higher-order functions |

First, we ignore the *do-while*. It doesn't exist in Scala.

As for the Scala *for* statement, it behaves like the Java *foreach* statement. That is, it operates on a collection and visits every element unconditionally. The Scala `for` statement *may* contain an *if* keyword. However, this case is covered by the *if* signature in Table 1.

We say the above signatures are “hard” in the sense that their signatures are part of the language. Furthermore, we hard-code them in a program table we call the “book” which Sclastic searches when it parses the input source.

D. Hard signatures

Scala also makes decisions in the context of higher-order functions that take Boolean-returning function objects. We call these higher-order functions, *predicate contexts*.

Consider Snippet 1. The function literal, `p => p % 2 == 0`, determines whether an element of the `List` collection is even. The problem is that we must search the source for all references to predicate contexts like `filter`. As we show, the Scala standard library has hundreds of such methods, the signatures of which we can put in the book with the hard signatures.

Doing so solves only part of the problem. It does not allow for the Scala standard library to incorporate new predicate contexts or refactor old ones. Furthermore, a programmer may extend the Scala standard library and add new predicate contexts or create new classes and predicate contexts that are independent of the Scala standard library.

Our solution to these problems was to make two passes over the input during the parser phase. During the first pass the parser identifies method definitions that are predicate contexts, that is, “soft” signatures, which the parser adds to the book. During the second pass, the parser queries the book to identify decision points, hard and soft. (In practice, there are in fact two books, a “hard” one, which is “hard-coded” into Sclastic, and a “soft” one, which is created dynamically and stored in a database. A configuration switch tells Sclastic to generate the soft one and stop or load the soft one and continue analyzing the source.)

E. Soft signature miss rate

The book may still be incomplete. Namely, decision points that reference predicate contexts, which are not in the portfolio of repositories, will not be in the book. One solution is to inflate the portfolio with repositories until the book is “closed,” namely, all references to predicate contexts are

contained in the book. We consider this approach definitive but impractical. The universe of Scala repositories is likely large and not necessarily completely hosted by GitHub.

We have chosen instead to model the potential severity of the problem by estimating the probability of a soft signature not being in the book when it is needed—the *soft signature miss rate*. First, we have the probability of declared imports that do not have corresponding package exports in the portfolio. This is the *package miss rate*. However, the soft signature miss rate is likely a fraction of the miss package rate since not every imported package contains predicate contexts. For instance, Java imports will not have predicate contexts. In general, the majority of Scala repositories in the portfolio do not contain predicate contexts. Thus, the soft signature miss rate is a joint probability, namely, the missing package rate times the probability that a package has predicate contexts, assuming the two are mathematically independent.

For a random sample of repositories, we model the soft signature miss rate, S , as follows:

$$S \approx k \times w \quad (3)$$

where k is the observed missing package rate and w is the observed fraction of repositories that contain packages with predicate contexts. We observe the parameters, k and w , using the law of succession [25] and frequency data extracted from the portfolio.

Finally, there are reasons we suggest to exclude the Scala compiler / standard library repository from the portfolio. However, our analysis always includes in the book soft signatures from the entire portfolio of repositories, that is, including the Scala repository.

V. EXPERIMENTAL DESIGN

In this section, we describe the experimental design and give summary statistics for the portfolio.

A. Data

We created a portfolio of all the Scala repositories that GitHub identified as trending. This term, “trending,” is GitHub terminology, which by GitHub’s definition means a repository that “the GitHub community is most excited about”. The important thing is that GitHub selects these repositories when we specify the language, “scala.” GitHub returns the respective “trending” repositories as hyperlinks on several web pages.

The portfolio starts as a collection of downloaded zip files which are inputs to Sclastic. The portfolio includes the Scala compiler / standard library which are written largely in Scala⁵; the Twitter, Inc., server and libraries⁶; several, large commercially inspired repositories such as Lift [26] and Akka⁷; and many smaller and lesser known repositories for computational finance, graphics, games, networking, web

⁵See “Scala: Object-Oriented Meets Functional,” <http://scala-lang.org>, accessed 6 Jun 2014

⁶See “Twitter is built on open source software,” <http://twitter.github.io/>, accessed 30 Jun 2013

⁷See “Akka,” <http://akka.io/>, accessed 6 June 2014

services, crypto-graphics, and artificial intelligence, among others.

In the case of Twitter, Inc., of its 42 repositories on Twitter's home page on GitHub⁶, three were found to be "trending" and the rest we included in the portfolio for the sake of curiosity and possible future research. The other exception was Casbah⁸, a repository we had used for initial testing.

We downloaded 262 repositories in total. 85% of the repositories were trending and the rest, 15%, were non-trending, being the 39 Twitter repositories and Casbah. This portfolio consisted of 21,596 source files with 2,391 KLOC (1,519 KLOC with comments and empty lines removed), and 223,493 methods.

The book has 1,187 soft signatures. 471 or approximately 40% are from the Scala repository. The remaining 60% are made from 77 other repositories. The portfolio exports 3,667 unique packages and imports 89,131 packages, 81,285 or 91.2% of which Sclastic found in the portfolio.

The table below gives statistics on the ten largest repositories in the portfolio ranked by number of methods. (All counts are $\times 1000$.)

TABLE II. TEN LARGEST REPOSITORIES IN THE PORTFOLIO.

| | Repository | Methods | % tot. | Raw LOC | Stripped LOC |
|----|--------------|---------|--------|---------|--------------|
| 1 | Scala | 57 | 26 | 401 | 247 |
| 2 | Scala Test | 17 | 7 | 300 | 170 |
| 3 | Delite | 9 | 4 | 62 | 41 |
| 4 | Lift | 9 | 4 | 106 | 58 |
| 5 | Akka | 8 | 4 | 105 | 66 |
| 6 | SBT-0.13 | 7 | 3 | 45 | 36 |
| 7 | Spire-2.10.0 | 5 | 2 | 23 | 17 |
| 8 | Scalaz-Seven | 5 | 2 | 42 | 28 |
| 9 | Finagle | 5 | 2 | 56 | 41 |
| 10 | BIDMat | 4 | 2 | 16 | 14 |

These ten largest repositories account for 56% of the methods and 47% of the executable LOC in the portfolio.

B. Setup

We used Eclipse⁹, Indigo service release 2 to develop and run Sclastic with the Scala 2.92 compiler and the Scala IDE plugin 3.0.0¹⁰.

We have one Korn shell script. It computes the package miss rate given a list of imports and a list of imports that have no declared class or package in the source. We also have one C program. It calculates, Kendall's τ [27] and MADM [28] statistics.

C. Nonparametric methods

A visual inspection of the distributions of M and LOC suggested the data probably were non-Gaussian. This was in

fact confirmed by the Kolmogorov-Smirnov (K-S) test [27]. Thus, we used only robust statistical measures like the K-S test. Two other methods we use are Kendall's τ and *median absolute deviation from the median* or MADM. Kendall's τ is a rank-based measure of correlation, a nonparametric analogue of Person's r . MADM is a rank-based measure of variability which might be considered a nonparametric analogue of the coefficient of variation. We calculate both of these statistics using the `kendall.c` program included in the Sclastic repository. The interested reader may wish to consult the source code and/or the literature for more details about these statistics.

D. Scatter plots

Our intentions for the scatterplots were to paint a picture of the qualitative relationship between M and LOC. However, since both M and LOC are discrete integer values, we found a simple scatterplot gives a terrace-like picture, obscuring many data points that may be overlaid by other data points. This loses much information. The scatterplots we use attempt to correct this problem by rendering data points at not at $x=M \times q$ and $y=LOC \times q$ but $x=\Omega(M \times q + \eta_0)$ and $y=\Omega(LOC \times q + \eta_1)$. Here q is a scaling constant that converts the respective value to pixel units (q is the same in both cases); η_0 and η_1 are uniform random deviates on the interval $[-0.50, 0.50]$; and Ω rounds to the nearest integer. In other words, we render each point without bias within one scaled unit of its location in the chart.

VI. RESULTS

Thus, per Equation 3 we estimate $k = 0.088$ and $w = (1+78) / (262+2)$, namely, in accordance with the law of succession [25]. Our "best guess" of the soft signature miss rate is $S \approx 0.026$.

All results are based on source after the comments and empty lines have been removed. Furthermore, since the Scala compiler/standard library repository is by far the largest in the portfolio, we analyzed the portfolio with this repository and without it to check for any possible bias the Scala repository may have had on the overall results. The table below gives the summary statistics for the portfolio with and without the Scala repository.

TABLE III. SUMMARY STATISTICS WITH AND WITHOUT THE SCALA REPOSITORY

| | Portfolio | w/o Scala repos. |
|----------------------|-----------|------------------|
| τ | 0.258 | 0.274 |
| Median M | 1.0 | 1.0 |
| Median LOC | 2.0 | 2.0 |
| MADM (M) | 0.0 | 0.0 |
| MADM (LOC) | 1.0 | 2.0 |
| Hard decision points | 126,432 | 96,732 |
| Soft decision points | 122,202 | 110,996 |

The scatter plots below include the Scala compiler/standard library repository and excludes it respectively.

⁸See "MongoDB," <http://10gen.com>, accessed 6 June 2014

⁹See "Eclipse," <http://eclipse.org>, accessed 15 Feb 2013

¹⁰See "Scala IDE for Eclipse," <http://scala-ide.org/>, accessed 12 Aug 2013

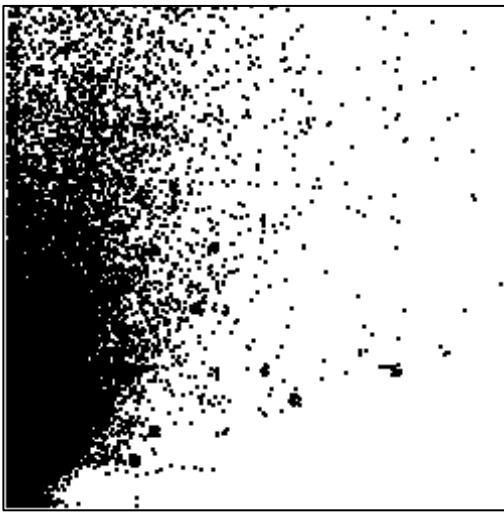


Fig. 1. Scatter plot M (horizontal axis) vs. LOC (vertical axis) including the Scala compiler/standard library repository. Both axes have range [0,50].

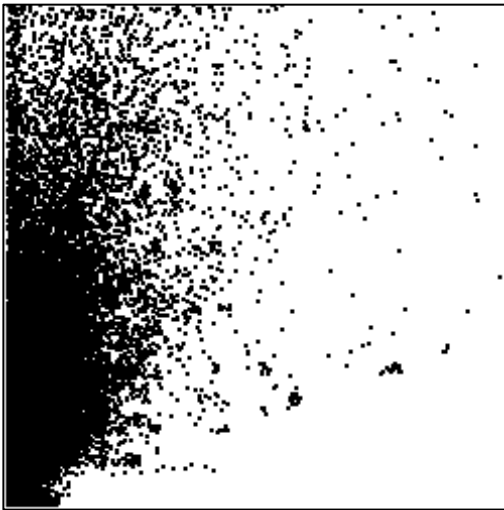


Fig. 2. Scatter plot M (horizontal axis) vs. LOC (vertical axis) excluding the Scala compiler/standard library repository. Both axes have range [0,50].

M is the horizontal axis and LOC, the vertical axis. The ranges of M and LOC on each axis are 0 - 50 (inclusive) which account for >99% of the data points.

The table below gives the distribution of the first ten M per method values across the entire portfolio.

TABLE IV. DISTRIBUTION OF M MEASUREMENTS

| M | Freq. | % of total | cum. % |
|-----|---------|------------|--------|
| 1 | 167,717 | 75.1 | 75.1 |
| 2 | 25,527 | 11.0 | 86.1 |
| 3 | 10,969 | 4.9 | 91.0 |
| 4 | 6,013 | 2.6 | 93.6 |
| 5 | 4,124 | 1.8 | 95.4 |
| 6 | 2,287 | 1.0 | 96.4 |
| 7 | 1,606 | 0.7 | 97.1 |
| 8 | 1,071 | 0.5 | 97.6 |
| 9 | 911 | 0.4 | 98.0 |
| 10 | 678 | 0.3 | 98.3 |

The chart below gives the M per method distribution plotted on a log-log scale.

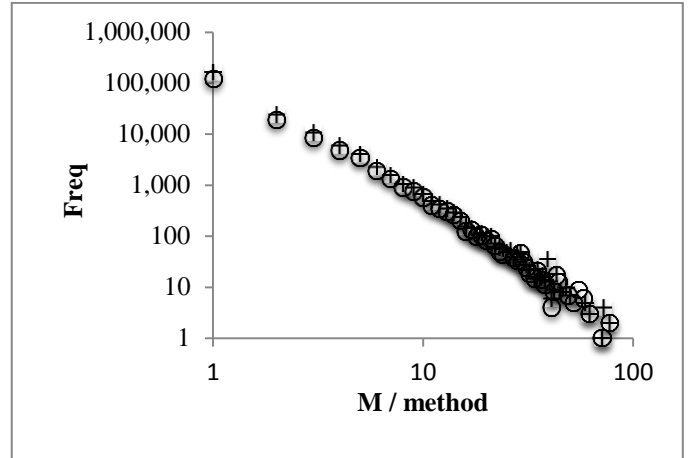


Fig. 3. Distribution of M / method plotted on log-log scales including (+) and excluding (o) the Scala compiler/standard library repository.

The table below gives the distribution of the first ten M per method values

TABLE V. DISTRIBUTION OF LOC MEASUREMENTS

| LOC | Freq. | % of total | cum. % |
|-----|---------|------------|--------|
| 1 | 100,692 | 45.5 | 45.5 |
| 2 | 31,221 | 14.1 | 58.6 |
| 3 | 14,212 | 8.7 | 68.3 |
| 4 | 9,994 | 6.4 | 74.7 |
| 5 | 7,197 | 4.5 | 79.2 |
| 6 | 2,287 | 3.3 | 82.5 |
| 7 | 1,606 | 2.7 | 85.2 |
| 8 | 5,990 | 2.1 | 87.3 |
| 9 | 4,700 | 1.7 | 89.0 |
| 10 | 3,935 | 1.5 | 90.5 |

The chart below gives the LOC per method distribution plotted on a log-log scale.

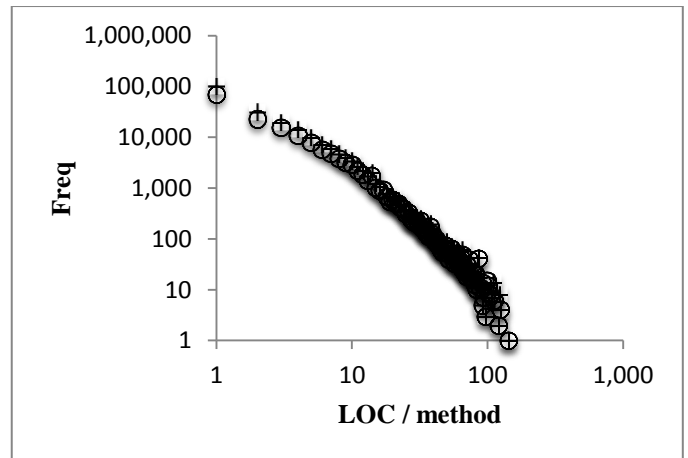


Fig. 4. Distribution of LOC / method plotted on log-log scales including (+) and excluding (o) the Scala compiler/standard library repository

VII. DISCUSSION

In this section we discuss and interpret the results.

A. Soft signature miss rate

We had noted that estimated soft signature miss rate, S , is less than 3%. This suggests that the book, which as we mentioned always contains the Scala standard library, is mostly complete as it contains an overwhelming majority of all the soft signatures required by the portfolio to reliably estimate M .

B. Correlations

Table 3 shows weak correlation between M and LOC including the Scala repository. The correlation also remains weak without the Scala repository. We interpret the weak correlation with and without Scala repository this to mean that the Scala repository does not appear to bias the M and LOC correlation. That the correlation is positive agrees with the commonsense expectation we posited at the outset. However, that the correlation is weak tell us that M is not “effectively useless” in relation to LOC as Hatton wrote.

The positive but weak correlation would seem to suggest that M and LOC are measuring related but nevertheless different phenomena in the source. Some evidence in support of this conjecture is the MADM statistics. That $MADM(M) < MADM(LOC)$ in general indicates that M is a more stable measure with less variability compared to LOC. This stands to reason since the range of M tends to be smaller than that for LOC. Indeed, this would explain the vertical layering of the scatterplots in Figure 1 and Figure 2. (Note: since $M > 0$ and $LOC > 0$, we find no data points on the $x=0$ or $y=0$ axis.) In other words, these data don’t contradict Hatton [23] but they also don’t fully support it. LOC and M are not interchangeable and both metric may be needed to provide a more complete picture of the complexity of Scala codes.

C. Hard and soft decision points

We note in Table 3 that there are nearly as many soft decision points as hard ones. The hard-to-soft ratio with the Scala repository is 1.03 and without it, 0.87. This fraction might indicate that overall programmers are exploiting the blend of functional and object styles in Scala, which would make sense. That the Scala repository employs fractionally more hard decision points (126,432-96,732=29,700) than soft ones (122,202-110,996=11,206) is noteworthy as the hard-to-soft ratio is 2.65. We offer only as conjecture the possibility that the Scala repository doesn’t reference its own standard library in relative terms. The standard library would be designed and implemented more for reuse by others.

D. Distributions

Although the median $M=1$ in Table 3, Table 4 shows that slightly more than 75% of methods have $M=1$. Although the median $LOC=2$, Table 5 shows nearly 70% of methods have $LOC \leq 3$. In other words, most of the code is highly modular and mostly simple. As we pointed out, the K-S test indicates that both of these distributions are non-Gaussian ($P < 0.01$).

In our opinion, this is perhaps even more interesting and surprising. First, on its face, this data tends to agree with

claims of functional programming proponents, that is, functional programming encourages highly modular coding. It does not, at least, seem to contradict them. Whether this is unique to Scala or the functional style is unknown. Second, it could be argued that the short and simple methods are mainly “getters” and “setters”. We don’t know; Sclastic does not distinguish getters and setters from other methods. However, we doubt this is the explanation for the preponderance of short, straight-line methods since Scala obviates the use of such boilerplate in general. Another explanation to consider is programmers are merely following the published style guides by Scala language designers and Twitter, Inc.¹¹ The problem with this idea is the style guides are only those: guides. Furthermore Scala is a relatively new language and the style guides, as far as we know, are even newer.

There is yet another possibility to account for these distributions. As we pointed out, the distributions and M and LOC are non-Gaussian. This was the reason we used robust methods of statistical analysis. First, the charts in Figure 3 and Figure 4 strongly resemble one another. Again, this suggests that with (+) or without (o) the Scala compiler / standard library repository, the general statistical pattern persists. Second, the distributions resemble those distributions of physical and aesthetic phenomena known to follow power-laws [29]. That is, the explanatory model has the form of a homogenous power-law, namely, $f(x) = c x^a$ where c and a are constants. This notion was tested by [30] which found power-laws offered the best, most parsimonious explanation for distributions and M and LOC. The reader will note that, indeed, if we plotted, $\log f(x) = a \log(x) + \beta$ we would obtain a line with slope a and intercept $\beta = \log(c)$. Figure 3 and Figure 4, in this case, $a < 0$, suggest that.

Here we wish to go further and speculate that the M and LOC type-distributions as presented in this paper may not be unique to Scala *per se*. Rather, they may be a statistical characteristic of other languages, when studied in the large as the case of our portfolio of Scala repositories. However, similar distributions for other languages have not been reported elsewhere in the literature, which leaves open a research for further study.

VIII. CONCLUSIONS

The results we have give in this paper point in a few different directions for future research. One of these is to confirm our findings for other functional programming languages where open source is concern. In this way, we have the opportunity to study possibly many other repositories. We gave a list of candidate languages in the “Background” section. Another direction is to study a language like Java. The promise of Java is we would likely find many repositories on GitHub. Finally, a study of Java repositories, being largely object-oriented at this time (Java 8, which supports lambda expressions, was released in March 2014), offers an opportunity to make some assessment and comparison of the relative contributions of functional and object styles in the data we presented here for Scala.

¹¹See “Scala Style Guide,” <http://docs.scala-lang.org/style/>, accessed 9 June 2014 and “Effective Scala,” <http://twitter.github.io/effectivescala/>, accessed 9 June 2014

ACKNOWLEDGMENTS

The authors thank the reviewers for their helpful feedback.

REFERENCES

- [1] M. Odersky, L. Spoon, B. Venners, *Programming in Scala: A Comprehensive Step-by-Step Guide*, Artima, 2011
- [2] M. Odersky, T. Rompf, "Unifying Functional and Object-Oriented Programming with Scala," *CACM*, vol. 57, no. 4, April 2014
- [3] T. McCabe, "A Complexity Measure," *IEEE Transactions on Software Engineering*, vol. SE-2, no. 4, December 1976
- [4] P. Hudak, "Conception, Evolution, and Application of Functional Programming Languages," *ACM Computing Surveys*, vol. 21, no. 3, 1989
- [5] G. O'Regan, *A Brief History of Computing*, Springer, 2010
- [6] C. Emerick, B. Carper, C. Grand, *Clojure Programming*, O'Reilly, 2012
- [7] M. Fogus, *Functional JavaScript: Introducing Functional Programming with Underscore.js*, O'Reilly, 2013
- [8] M.R. Hsen, H. Rischel, *Functional Programming Using F#*, Cambridge University Press, 2013
- [9] G. Michaelson, *An Introduction to Functional Programming Through Lambda Calculus*, Dover, 2011
- [10] S. St. Laurent, *Introducing Erlang*, O'Reilly, 2013
- [11] S. Thompson, *Haskell: The Craft of Functional Programming*, Pearson Education Ltd, 2011
- [12] D. Wampler, *Functional Programming for Java Developers: Tools for Better Concurrency, Abstraction, and Agility*, O'Reilly, 2011
- [13] H. Sutter, "The Free Lunch is Over: The Fundamental Turn Toward Concurrency in Software," *Dr. Dobbs Journal*, vol. 30, no. 3, 2005
- [14] R. Coleman, U. Ghattamaneni, "Parallel Collections: A Free Lunch?," *Journal of Computer Science and Engineering*, vol. 17, issue 2, 2012
- [15] J. Hughes, "Why Functional Programming Matters," *Research Topics in Functional Programming*, ed. Turner D, Addison-Wesley, 1990, pp. 17–42
- [16] I. Kant, *The Critique of Judgment* (1790), translation by J. C. Meredith, Oxford University Press, 1978
- [17] E. Weyuker, "Evaluating Software Complexity," *IEEE Transactions on Software Engineering*, vol. 14, no. 9, Sep. 1988
- [18] D. Tran-Cao, G. Lévesque, J. Meunier. "A Field Study of Software Functional Complexity Measurement." *Proceedings of the 14th International Workshop on Software Measurement*, 2004
- [19] G. Gill, C. Kemerer C, "Cyclomatic Complexity Metrics Revisited: An Empirical Study of Software Development and Maintenance," *CISR WP No. 218, Sloan WP No. 3222-90*, 1990
- [20] N. Pataki, A. Sipos, Z. Porkolab, "Measuring the Complexity of Aspect-Oriented Programs with Multiparadigm Metric," *Proc. of ECOOP 2006 Doctoral Symposium and PhD Students Workshop*, 2006
- [21] SEI, "C4 Technology Reference Guide, Software Engineering Institute," Carnegie Mellon, 1997
- [22] C. Archer "Measuring Object-Oriented Software Products," *Software Engineering Institute, Carnegie Mellon*, 1995
- [23] L. Hatton, "The role of empiricism in improving the reliability of future software," *TAIC*, 2008
- [24] N. Nystrom, W. White, K. Das, "Firepile: GPU Programming in Scala," *GPCE*, 23 Oct 2011
- [25] E.T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge, UK, Cambridge University Press, 2003
- [26] T. Perrett, *Lift in Action: The Simply Functional Web Framework for Scala*, Manning Publications, 2011
- [27] J. Conover J, *Practical Non-Parametric Statistics*, Wiley, 1995
- [28] D.C. Hoaglin, F. Mosteller, J.W. Tukey, *Understanding Robust and Exploratory Data Analysis*, Wiley-Interscience, 2000
- [29] M. Schroeder, *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*, Dover, 2009
- [30] R. Coleman, M. Johnson, "Power-Laws and Structure in Functional Programs", *Proceedings of the 2014 International Conference on Computational Science & Computational Intelligence*, Las Vegas, NV, 10 – 13 Mar, 2014, IEEE Computer Society CPS.

A Crypto-Steganography: A Survey

Md. Khalid Imam Rahmani¹

¹Associate Professor, Deptt. of Computer Sc. & Engg.
Echelon Institute of Technology
Faridabad, INDIA.

Kamiya Arora², Naina Pal³

^{2,3}M.Tech. Scholar, Deptt. of Computer Sc. & Engg.
Echelon Institute of Technology
Faridabad, INDIA.

Abstract—The two important aspects of security that deal with transmitting information or data over some medium like Internet are steganography and cryptography. Steganography deals with hiding the presence of a message and cryptography deals with hiding the contents of a message. Both of them are used to ensure security. But none of them can simply fulfill the basic requirements of security i.e. the features such as robustness, undetectability and capacity etc. So a new method based on the combination of both cryptography and steganography known as Crypto-Steganography which overcome each other's weaknesses and make difficult for the intruders to attack or steal sensitive information is being proposed. This paper also describes the basics concepts of steganography and cryptography on the basis of previous literatures available on the topic.

Keywords—Steganography; Image Steganography; Cryptography; Least Significant Bit (LSB); Enhanced Least Significant Bit (ELSB); Compression; Decompression; Advanced Encryption Standard (AES); Data Encryption Standard (DES); Hashing algorithms

I. INTRODUCTION

It's a well-known fact that security of data has become a major concern nowadays. The growth of modern communication technologies imposes a special means of security mechanisms especially in case of data networks [33]. The network security is becoming more important as the volume of data being exchanged over the Internet increases day by day [29].

The two important techniques for providing security are cryptography and steganography [5]. Both are well known and widely used methods in information security.

One of the reasons why the attackers become successful in intrusion is that they have an opportunity to read and comprehend most of the information from the system [29]. Intruders may reveal the information to others, misuse or modify the information, misrepresent them to an individual/organization or use them to plan even some more severe attacks [13]. One of the solutions to this problem is through the use of steganography and cryptography.

Steganography is the art of hiding information in digital media through the techniques of embedding hidden messages in such a way that no one except the sender and the intended receiver(s) can detect the existence of the messages [1].

Cryptography is the art of transmitting the data safely over the Internet by applying some cryptographic algorithms so that it will be difficult for an attacker to attack or steal some confidential or private information [11].

A brief description of the state of the art of steganography and cryptography is given in section II. A literature survey is described in section III. In section IV, a proposed algorithm has been described. Section V describes the comparative analysis of both the techniques. Section VI has been used for the conclusion and future direction of the research work.

II. THE STATE OF THE ART

A. Steganography

Steganography is the technique of embedding hidden messages in such a way that no one, except the sender and intended receiver(s) can detect the existence of the messages. The main goal of steganography is to hide the secret message or information in such a way that eavesdroppers are not able to detect it [1]. If they found any suspicious data, then goal is defeated. Other goal of steganography is to communicate securely in a completely undetectable manner. The various forms of data in steganography can be audio, video, text and images etc. The basic model of Steganography consists of three components [3]:

- The Carrier image: The carrier image is also called the cover object that will carry the message that is to be hidden.
- The Message: A message can be anything like data, file or image etc.
- The Key: A key is used to decode/decipher/discover the hidden message.

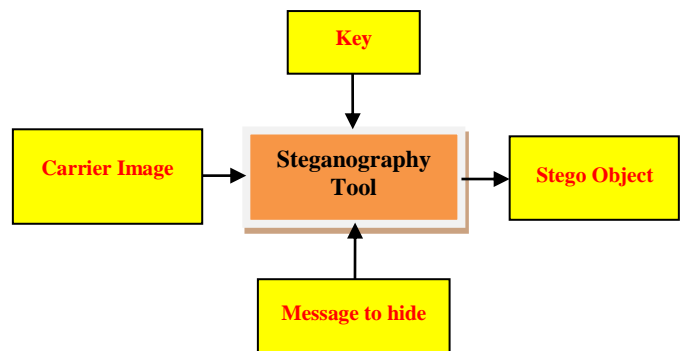


Fig. 1. Basic Model of Steganography

The encryption and decryption processes for hiding an image in steganography can be defined in a flow chart [3] as below:

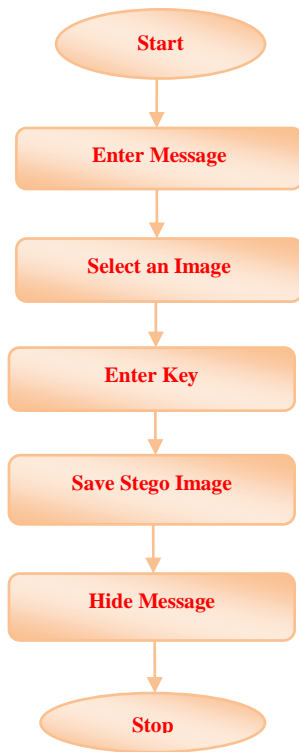


Fig. 2. Embedding Secret Message into Cover Object

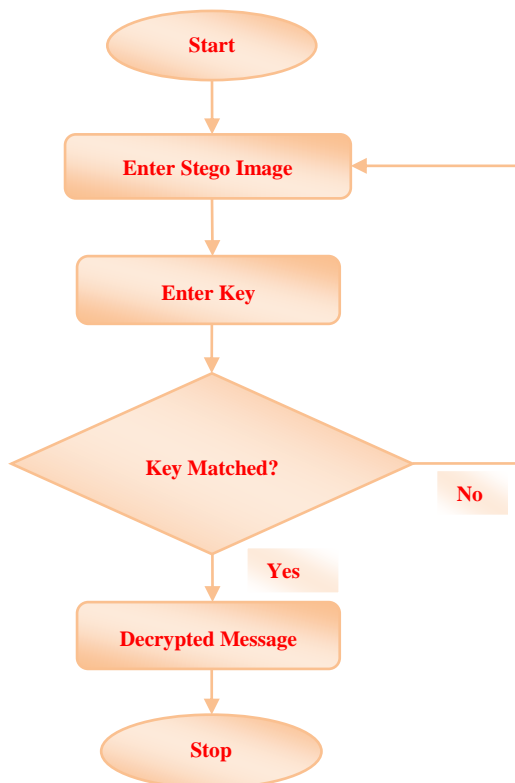


Fig. 3. Extracting Secret Message from Cover Object

1) Types of Steganography

The various types of steganography include [17]:

a) Image Steganography

The image steganography is the process in which we hide the data within an image so that there will not be any perceivable change in the original image. The conventional image steganography algorithm is LSB embedding algorithm.

b) Audio Steganography

The method of hiding secret information in an audio is known as audio steganography. There are various methods for hiding secret data in an audio such as LSB, Phase Coding etc. [17].

c) Video Steganography

The method of hiding secret information in a video is known as video steganography. Video consist of images as well as audio. Hence, both images and audio steganography can be used for video steganography [17].

d) Text files Steganography

The method of hiding secret information in a text is known as text steganography. Text steganography requires less memory as it can only store text files. It provides quick communication or transfer of files from one computer to another. Text steganography is not commonly used as text files containing large amount of redundant data [17].

2) Techniques of Steganography

Steganography techniques are as follows:

a) Least Significant Bit (LSB)

Least Significant bit is the most common technique used for hiding the secret information in any digital media like image, text or audio/video. LSB refers to replacement of last bit of an image with the bit of secret message [23]. One can use 8 bit or 24 bit image to hide data. 24 bit images are well suited for hiding large amount of data. Although LSB is simple and useful for the user but it can be detected by an attacker during transmission of data on the network. There are many versions of LSB like Edge-LSB, Random-LSB and Enhanced LSB etc. [23].

b) Bitmap Steganography

There are two types of compressions: Lossy compression and Lossless Compression. In lossy compression, the data can be lost after applying compression while in lossless compression, the data can't be lost. For compression of images, lossy compression is generally used wherein after the compression the image can be restored but its quality can be degraded [21]. Bitmap steganography is the simple and most common approach as only BMP files gives lossless compression. BMP images are created from pixels and all pixels are comprised of three basic components i.e. Red, Green and Blue and named as RGB.

A combination of these three color components can form every color that is seen in these images. It is known that every byte in Computer Science is created from 8 bits and the first

bit is called the most significant bit (MSB) and the last bit is called the least significant bit (LSB).

Suppose that there are three adjacent pixels (9 bytes) with the RGB encoding:

```
10010101 00001101 11001001
10010110 00001111 11001011
10011111 00010000 11001011
```

The decimal number 300 can be converted into binary representation which is 100101100. This representation can be embedded into the least significant bits of the image. LSB can be represented as (where bits in different color have been changed)

```
10010101 00001100 11001000
10010111 00001110 11001011
10011111 00010000 11001010
```

Here the number 300 was embedded into the grid, only the 5 bits are needed to be changed according to the embedded message. On an average, only half of the bits in an image would be modified to hide a secret message using the maximum cover size [32].

B. Cryptography

Cryptography is the art of achieving security by encoding messages to make them non-readable [34]. Cryptography is an art of transmitting the data safely over the Internet by applying some cryptographic algorithms so that it will be difficult for an attacker to attack or steal some confidential or private information.

Two basic terms used in cryptography are encryption and decryption; encryption is the process of converting plain text into cipher text and decryption is the reverse process of encryption [34]. Plain text is the text having the actual message or data which is not encrypted and cipher text is the text after encryption of message or data which is ready to be shared [34]. A key is needed for both encryption and decryption of the message.

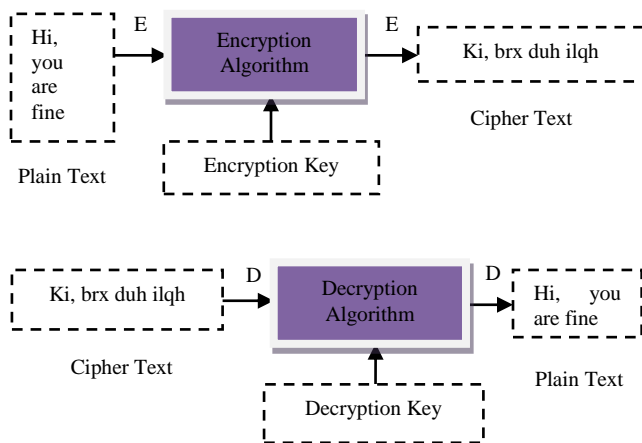


Fig. 4. Basic Model of Cryptography

C. Steganography vs Cryptography

TABLE I. DIFFERENCE BETWEEN STEGANOGRAPHY AND CRYPTOGRAPHY

| Techniques | Steganography | Cryptography |
|---------------------------|---|---|
| Definition | Steganography means cover writing | Cryptography means secret writing |
| Objective | Focuses on keeping existence of a message secret | Focuses on keeping contents of a message secret |
| Key | Optional | Necessary |
| Carrier | Any digital media | Usually text based |
| Visibility | Never | Always |
| Security Services Offered | Confidentiality, authentication | Confidentiality, availability, data integrity, non-repudiation |
| Attacks | It is broken when attacker detects that steganography has been used known as Steganalysis | It is broken when attacker can read the secret message known as Cryptanalysis |
| Result | Stego file | Cipher text |

III. LITERATURE SURVEY

In [1], authors explore the steganography, its history, features, tools and various techniques like LSB, masking, filtering and other transformations used for hiding messages in an image.

In [2], basic cryptographic concepts and techniques are defined. The paper also describes various methods to hide the secret or confidential message in an original file so that it is unintelligible to an interceptor.

In [3], Dr. R. Sridevi, Vijaya Lakshmi Paruchuri, K. S Sadasiva Rao introduced the concept of embedding the secret message into an image using LSB technique and then applied AES algorithm to provide better security.

The paper [4] proposes a reverse procedure described in paper [3] by using an alteration component method.

In [5] user enters username, password and a key. A key is taken from automatic key generator device which generates a unique key after some specific time. After this the secret message and key is encrypted and encrypted message is embedded into cover image and stego image is produced.

In paper [6] the secret message is first compressed then the message is hashed and encrypted using encryption key. This method results in robust model and achieves two important principles of security i.e. privacy and authenticity.

In [7], various technologies used in image steganography are proposed. This paper presents a review used for hiding a secret message or image in spatial and transform domain. This paper also proposed techniques for detecting the secret message or image i.e. steganalysis.

The paper at [8] introduced a method where secret message is first compressed using wavelet transform technique and then embeds into cover image using LSB where the bits of secret message is inserted into image by using random number generator.

In [9], A. Joseph Raphael introduces basic terminologies of cryptography and steganography and ensures that the combination of both gives multiple layers of security and will achieve requirements like capacity, security and robustness.

The paper at [10] introduced a method based on image ranking. Firstly, secret data is encrypted using RSA encryption algorithm and then users selects any image suited for hiding particular data. This will make difficult for attacker to succeed an attack. Finally, a stego image has been produced but this paper lacks in integrity and this application cannot hide large data.

In [11], authors give brief review of above techniques used for ensuring security. It is proved in this paper that using these techniques, data can be made more secure and robust.

The authors in paper [12] introduced the method for embedding the secret image into cover image using LSB technique and then encrypts using DES algorithm and used the key image.

In [13], authors first embed the secret data within cover image using LSB technique and then apply DES encryption method for encrypting the data which provides better security.

In [14], authors first encrypts the data with RC4 encryption algorithm and then embeds in BMP cover image using three different steganographic methods and then compares these three methods. This paper also results in achieving the requirements of security i.e. data confidentiality, data integrity and data authentication.

The paper at [15], embeds the secret image into 24 bit or 8 bit image by using LSB and then evaluated results for 2, 4, 6 LSB for a .png file and a .bmp file.

In [16], authors proposed a new technique called metamorphic cryptography where secret image is encrypted and transformed into a cipher image using key and this cipher image is embedded into a cover image by converting it into an intermediate text and finally transformed once again into an image.

In the paper at [17], authors define basic terminologies of steganography, steganography techniques, classifications and review of previous work done by researchers.

In the paper at [18], authors define a method of hiding information on the billboard. This method can be used for announcing a secret message in public place. User first enters the normal data then hides secret data into normal data and the encrypted data is displayed on the billboard board. This encrypted data is saved for decrypting the secret data.

In paper [19], user selects secret image in BMP format and encrypts using BLOWFISH cryptography Algorithm because BLOWFISH is faster, stronger and gives good performance when compared with DES, 3DES, AES, RC6, RC4. This

encrypted image is embedded into video using LSB technique and forms stego video. This method provides confidentiality, authenticity, integrity and non-repudiation.

In [20], authors used a different approach to hide an image i.e. Hide behind Corner (HBC) algorithm is used to place a key at the image corners. All the keys at the corners are encrypted by generating Pseudo Random Numbers. Then the hidden image is transmitted. The receiver should know all the keys that are used at the corners while encrypting the image. Reverse Data Hiding (RDH) is used to get the original image and the original image is produced when all the corners are unlocked with proper secret keys used for hiding the image.

In [21], user enters username, password to login into the system. After successful login, user can embed secret message into an image using a key and produces stego image. Same key is used at receiver site for retrieving the hidden data. Here the secret message is transferred into text file first. Then the text file is compressed into zip file, the zip text file then is used for converting it into binary codes. Zipping the text file is more secured and is hard to detect.

In [22], authors present a new technique for hiding information based on Huffman encoding. The gray level image of size $m*n$ and $p*q$ is taken as cover image and secret image respectively. The Huffman encoding is performed over secret image and each bit of Huffman code of secret image or a message is embedded into cover image by using LSB.

The paper at [23] is similar to paper at [10] where secret data is encrypted using RSA encryption algorithm and then user selects any image suited for hiding particular data and then this secret data is embedded into cover image using LSB. This will make difficult for an attacker to steal sensitive information. Finally, a stego image has been produced.

In [24], paper presents a method for encrypting and decrypting a secret file which embeds into image file using random LSB insertion method in which bits of secret message are spread into image bits randomly. These random numbers are generated by using a key.

In [25], the secret message or data can be hidden in any image, audio or video which provides more security. The secret data is first encrypted using AES algorithm and key is hashed using SHA-1 to prevent from attacks then user can hide the cipher data in image, audio or video using LSB technique. The receiver should provide the same key that is hashed for encryption.

The paper [26] is similar to paper [19] but the only difference is that here user selects plain text and encrypts using BLOWFISH Algorithm. This encrypted text is embedded into image using LSB technique and forms stego image. Reverse procedure is done for decrypting the secret image.

In [27], the method is similar to that in the paper [6]; the only difference is the compressed message is not hashed. This novel approach requires less memory space, fast transmission rate, better security and no distortion in quality of image.

In the paper at [28], authors proposed two methods to ensure high security. First method includes the combination of

both steganography technique and cryptography technique and second method only includes steganography approach.

IV. THE PROPOSED ALGORITHM

Based on the findings in the existing papers studied, a new algorithm is being proposed that can ensure all of the security principles i.e. robustness, confidentiality, authentication, integrity and non-repudiation and that would also satisfy the requirements of steganography i.e. capacity, undetectability and robustness [18]. The algorithm which will be implemented in a proposed system at a later stage of this research work consists of four layers. Each layer is used to achieve one security principle like layer 1 implements authorization, layer 2 implements authentication, integrity and non-repudiation, layer 3 implements confidentiality and partial security, layer 4 implements robustness and the remaining part of security. Each layer defined is unperceivable for an attacker. The flow chart below describes various steps involved in the algorithm:

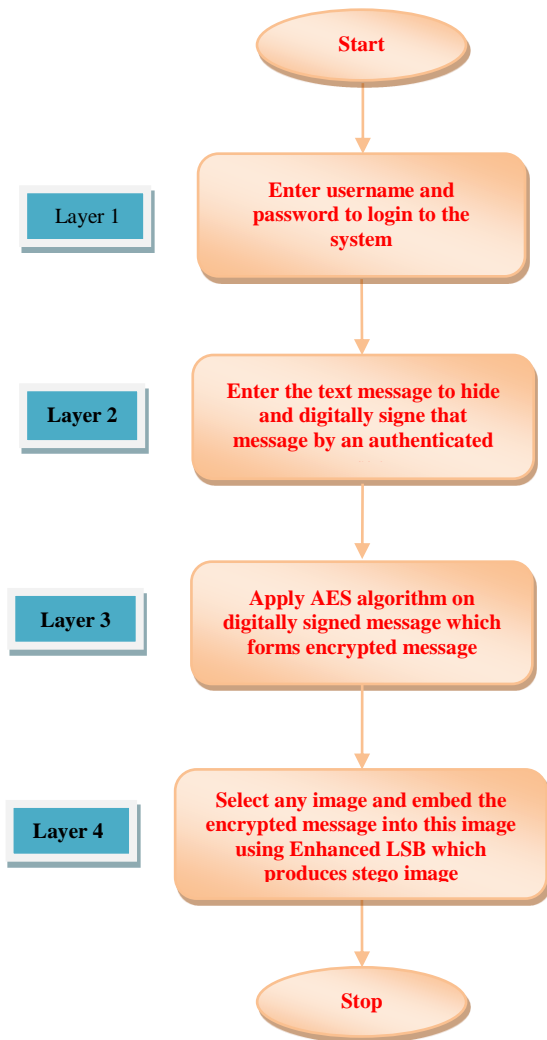


Fig. 5. Proposed Algorithm

V. COMPARATIVE ANALYSIS

A comparative analysis is made to compare the effectiveness of the proposed method with respect to other available methods. The comparative analysis has been performed on the basis of the requirements of security i.e. confidentiality, robustness and authentication etc. From this analysis, it has been identified that all the papers discussed in literature review lack in some aspect or the other as far as the implementation of the principles of security is concerned.

The effectiveness of the proposed method can be estimated by computing some valuable statistical operations.

TABLE II. COMPARATIVE ANALYSIS OF THE LITERATURE

| Literature Reference | Requirements | | |
|----------------------|-----------------|------------|----------------|
| | Confidentiality | Robustness | Authentication |
| [2] | Yes | Yes | No |
| [3] | Yes | No | No |
| [4] | Yes | No | No |
| [5] | No | No | Yes |
| [6] | Yes | No | Yes |
| [7] | Yes | Yes | No |
| [9] | Yes | Yes | Yes |
| [10] | No | No | Yes |
| [13] | Yes/No | No | No |
| [14] | Yes | Yes | Yes |
| [15] | Yes/No | No | No |
| [16] | Yes | Yes | No |
| [18] | No | No | No |
| [20] | Yes | Yes | No |
| [21] | Yes | Yes | Yes |
| [22] | Yes | Yes | No |
| [24] | Yes | Yes | No |
| [25] | Yes | Yes | No |
| [27] | No | No | No |

VI. CONCLUSIONS

In this paper, a very comprehensive review of the conventional approaches and techniques used in the security of transmitted data over the data networks has been given. The survey has been carried out related to both steganography and cryptography that ensures security but lacks in some way or the other as far as their individual capabilities related to coverage of all the security principles are concerned. So, in order to overcome the lack of coverage of all the principles of security in those algorithms, a new algorithm has been proposed that would satisfy all the principles of security and also satisfy the requirements of steganography.

The proposed algorithm can be implemented in a security system as a future research work that would probably excel in comparison to the existing algorithms. The system would be tested on the basis of various test cases and the results would be compared with those of existing algorithms.

REFERENCES

- [1] Neil F. Johnson, Sushil Jajodia, "Exploring Steganography: Seeing the Unseen", IEEE, Feb1998, pp. 26-34.
- [2] F. Piper, "Basic Principles of Cryptography", IEEE Colloquium on Public uses of Cryptography, April 1996, pp. 2/1-2/3.
- [3] Dr. R. Sridevi, Vijaya Lakshmi Paruchuri, K.S. Sadasiva Rao, "Image Steganography combined with Cryptography", International Journal of Computers & Technology, ISSN: 22773061, Vol.9, July 2013, pp. 976-984.
- [4] Lokesh Kumar, "Novel Security Scheme for Image Steganography using Cryptography Technique", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X Vol.2, April 2012, pp. 143-146.
- [5] Mihir H Rajyaguru, "Cryptography -Combination of Cryptography and Steganography with Rapidly Changing Keys", International Journal of Emerging Technology and Advanced Engineering, ISSN: 2250-2459, Vol.2, October 2012, pp. 329-332.
- [6] H.Al-Barhmtoshy, E.Osman and M.Ezzaand, "A Novel Security Model Combining Cryptography and Steganography", Technical Report, 2004, pp. 483-490.
- [7] S.Ashwin, J.Ramesh, K.Gunavathi, "Novel and Secure Encoding and Hiding Techniques Using Image Steganography: A Survey", IEEE Xplore International Conference on Emerging Trends in Electrical Engineering and Energy Management, Dec 2012, pp. 171-177.
- [8] Humanth Kumar, M.Shareef, R. P. Kumar, "Securing Information Using Steganography", IEEE Xplore International Conference on Circuits, Pwer and Computing Technologies, March 2013, pp. 1197-1200.
- [9] A. Joseph Raphael, Dr. V.Sundaram, "Cryptography and Steganography-A Survey, International Journal of Computer and Technology Applications", ISSN: 2229-6093, Vol.2 (3), 2010, pp. 626-630.
- [10] Armin Bahramshahry, Hesam Ghasemi, Anish Mitra, Vinayak Morada, "Design of a Data Hiding Application Using Steganography", Databases, 2007, pp. 1-6.
- [11] Vipula Madhukar Wajgade, Dr. Suresh Kumar, "Stegocrypto - A Review of Steganography Techniques using Cryptography", International Journal of Computer Science & Engineering Technology, ISSN: 2229-3345, Vol. 4, 2013, pp. 423-426.
- [12] R.Nivedhitha, Dr.T.Meyyappan, "Image Security using Steganography and Cryptographic Techniques", International Journal of Engineering Trends and Technology, ISSN: 2231-5381, Vol.7, 2012, pp. 366-371.
- [13] Dhawal Seth, L. Ramanathan, Abhishek Pandey, "Security Enhancement: Combining Cryptography and Steganography", International Journal of Computers Applications, ISSN: 0975-8887, Vol. 9(11), 2010, pp. 3-6.
- [14] Wai Wai Zin, "Implementation and Analysis of Three Steganographic Approaches", IEEE Xplore International Conference on Computer Research and Development, March 2011, pp. 456-460.
- [15] D. Jacobs, Snehal Kamalapur, Neeta Sonawane, "Implementation of LSB Steganography and its Evaluation for Various Bits", IEEE Xplore International Conference on Digital Information Management, Dec 2006, pp. 173-178.
- [16] N.V Rao, J.TL Philjon, "Metamorphic Crypto- A Paradox between Cryptography and Steganography using Dynamic Encryption", IEEE Xplore International Conference on Recent Trends in Information Technology, June 2011, pp. 217-222.
- [17] Mehdi Hussain, Mureed Hussain, "A Survey of Image Steganography Technique", International Journal of Advanced Science and Technology, Vol. 54, 2013, pp. 113-124.
- [18] S. Channalli, A. Jadhav, "Steganography an Art of Hiding Data", International Journal on Computer Science and Engineering, ISSN: 0975-3397, Vol.1(3), 2009, pp. 137-141.
- [19] Ms. Hemlata Sharma,Ms. MithleshArya, Mr. Dinesh Goyal , "Secure Image Hiding Algorithm using Cryptography and Steganography", IOSR Journal of Computer Engineering (IOSR-JCE), ISSN: 2278-8727, Vol. 13(5), August 2013, pp. 1-6.
- [20] Hemalatha M., Prasanna A., Dinesh Kumar R., Vinoth kumar D., "Image Steganography using HBC and RDH Technique", International Journal of Computer Applications Technology and Research, Vol.3, 2014, pp. 136-139.
- [21] Rosziati Ibrahim, Teoh Suk Kuan, "Steganography Algorithm to Hide Secret Message inside an Image", Computer Technology and Application, 2011, pp. 102-108.
- [22] Rig Das, Themrichon Tuithung, "A Novel Steganography Method for Image Based on Huffman Encoding", IEEE, 2012.
- [23] M.Juneja, P.S. Sandhu, "Data Hiding with Enhanced LSB Steganography and Cryptography for RGB Color Images", International Journal of Applied Research , ISSN: 2249-555X , Vol. 3(5), May 2013, pp. 118-120.
- [24] M.S Sutaone., M.V. Khandare, "Image Based Steganography using LSB Insertion Technique", IEEE Xplore, Jan 2008, pp. 146-151.
- [25] Shery Elizabeth Thomas, Sumod Tom Philip, Sumaya Nazar, Ashams Mathew, Niya Joseph, "Advanced Cryptographic Steganography using Multimedia Files", International Conference on Electrical Engineering and Computer Science (ICEECS), May 2012, pp. 239-242.
- [26] Ajit Singh, Swati Malik, "Securing Data by using Cryptography with Steganography", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Vol. 3(5), May 2013, pp. 404-409.
- [27] M. Sitaram Prasad, S. Nagan Janeyulu, Ch. Gopi Krishna, C. Nagaraju, "A Novel Information Hiding Technique for Security by using Image Steganography", Journal of Theoretical and Applied Information Technology, 2005-2009, pp. 35-39.
- [28] Khalil Challita, Hikmat Farhat, "Combining Steganography and Cryptography: New Directions", International Journal on New Computer Architectures and their Applications (IJNCAA), ISSN: 2220-9085, Vol. 1(1), 2011, pp. 199-208.
- [29] Jidagam Venkata Karthik, B.Venkateshwar Reddy, "Authentication of Secret Information in Image Steganography", International Journal of Latest Trends in Engineering & Technology, ISSN: 2278-621X, Vol. 3(1), Sep 2013, pp. 97-104.
- [30] Vipula M.Wajgade, Nagesh D. Matharia, Dr. Suresh Kumar, "Enhancing Data Security with Advanced Digital Image Steganography", International Journal of Pure and Applied Research in Engineering and Technology, ISSN: 2319-507X, Vol. 1(8), 2013, pp. 228-238.
- [31] M. Pavani, S. Naganjaneyulu, C. Nagaraju, "A Survey on LSB Based Steganography Methods", International Journal of Engineering and Computer Science (IECS), ISSN:2319-7242, Vol. 2 (8), August, 2013, pp. 2464-2467.

- [32] Shilpa Gupta, Geeta Gujral, Neha Aggarwal, "Enhanced Least Significant Bit algorithm for Image Steganography", International Journal of Computational Engineering & Management (IJCEM), ISSN: 2230-7893, Vol. 15(4), July 2012, pp. 40-42.
- [33] Aprajita, Ajay Rana, "Steganography-The Art of Hiding Information-Comparison from Cryptography", International Journal of Innovative Research in Science, Engineering and Technology, ISSN : 2319-8753, Vol. 1(5), May 2013, pp. 1308-1312.
- [34] Atul Kahate, "Cryptography and Network Security", Tata McGraw-Hill, 2006.

AUTHORS' BIBLIOGRAPHY



Md. Khalid Imam Rahmani is an Associate Professor in the Department of Computer Science & Engg. of a very reputed NBA accredited Engineering College, Echelon Institute of Technology, Faridabad, India. He is having about 17 years of teaching, industry and administrative experience. He has done B.Sc. Engg. in Computer Engineering from A.M.U., Aligarh, M.Tech. in Computer Engineering from M.D.U., Rohtak and is pursuing Ph.D. in Digital

Image Retrieval Algorithms. Digital Image Processing, Innovative Programming techniques, Mobile Computing, Algorithms Design and Internet & Web Technologies are his research areas.



Kamiya Arora has earned her M.Tech. degree in Computer Science & Engg. from Echelon Institute of Technology under Maharshi Dayanand University, Rohtak. Her research interests include Image Processing, Steganography, Data mining and Cryptography.



Naina Pal has earned her M.Tech. degree in Computer Science & Engg. from Echelon Institute of Technology under Maharshi Dayanand University, Rohtak. Her research interests include Image Processing, Classification of data using Clustering and Data mining.

An Ecn Approach to Congestion Control Mechanisms in Mobile Adhoc Networks

Som Kant Tiwari
Department of CSE
RITS, RGI,Group,
Bhopal,India

Dr. Y.K.Rana
Department of CSE
RITS, RGI,Group,
Bhopal,India

Prof. Anurag Jain
Department of CSE
RITS, RGI,Group,
Bhopal,India

Abstract—Node(s)/link(s) of a network are subjected to overloading; network performance deteriorates substantially due to network congestion. Network congestion can be mitigated with the help of Explicit Congestion notification (ECN) technique. ECN notification is carried out by setting ECN bit in the TCP header. This allows for end-to-end notification of network congestion without dropping packets. ECN bit notifies TCP sources of incipient congestion before losses occur. ECN is a binary indicator (1 bit) which does not reflect the congestion level completely and so its convergence speed is relatively low. In our work, we have used an extra ECN bit (2 bit ECN). The extra bit allows for passing of additional congestion feedback to the source node. This enables the source node to determine the level of congestion based on which steps can be taken to ensure faster convergence. In comparison to single bit ECN, the additional information afforded by double bit ECN allows for more flexibility to adjust window size, to handle congestion. Simulation results have shown that the proposed method improves overall performance of the network by over 12%.

Keywords—Explicit Congestion Notification (ECN); Mobile ad hoc Networks (MANET); Congestion control; Congestion window Transmission Control Protocol (TCP)

I. INTRODUCTION

Mobile Ad hoc Network (MANET) is a distributed system of self-organizing and independent nodes. Mobility of nodes causes Network topology to change frequently. In the absence of infrastructure (due to its Ad-hoc characteristics) and dynamic topology, MANET nodes must self-reorganize themselves to forward each other's data to achieve end-to-end communication. As such, MANET nodes must not only behave as end nodes, but also as routers for forwarding packets. The ability to re-organize and self-operate makes MANET suitable for a number of applications requiring ad-hoc operation, including- for military and emergency scenarios. Self re-organization and self-operation requires adapting to variations in topology and bandwidth requirements, which is the main challenge in the design of MANET. Each node stores data to be forwarded in a buffer, until it is successfully forwarded. Frequent changes in bandwidth requirement (load) and topology can cause the buffer to get full as new packets arrive before older packets are forwarded. This results in newly arriving packets not finding space in the buffer, to wait for their turn to be forwarded; leading to them being dropped (loss). In an attempt to adapt to topology and traffic dynamics, additional control packet traffic is generated. This, in addition to the attempts to

re-transmit lost packets, cause traffic surge at various nodes along the path. Nodes that is unable to cope with this surge form bottleneck. Bottleneck leads to congestion, which reduces the performance of whole network.

Mitigation of congestion (congestion control) is carried out at transport layer. In TCP (Transport Control Protocol) numbers of variants have been proposed for congestion control. They can be classified into two categories according to the mechanisms used to obtain congestion feedback: explicit network feedback based schemes and pure end-to-end schemes. Explicit network feedback based schemes usually perform better than pure end-to-end ones, but many of them require modifying IP header to carry additional feedback information. This incurs complicated computation in routers and makes them unpractical in real networks. For example, the XCP protocol can obtain accurate feedback, achieving high efficiency, fairness and fast convergence speed; but their additional feedback overhead also makes it difficult to deploy them in real networks. We propose an enhanced version of ECN scheme allows more flexible handling of congestion. This scheme is based on the network load factor ρ_l which can be estimated very easily at each router and is associated with following illustrated ECN bits in the IP header for feedback. We are making the sender aware about the state of congestion of link in the network. This is done in our scheme through explicit feedback from the network in the form of explicit congestion notification (ECN). Depending on the feedback from the network, the sender act accordingly sudden increase/ decrease the congestion window. By using two ECN bits, it gives more convergence to adjust window size with added options to RED queue to handle congestion more effectively.

This paper is organized as follows: existing work described in section II, followed by proposed methodology is described in section III, simulation and result present in section IV, finally, we conclude this work in section V.

II. EXISTING WORK

The new queue management scheme is designed [8] to help the monitoring the global congestion situation of an autonomous system. In order to observe the congestion situation of the system, traffic is generated between routers and a centralized unit. Routers are send packets according to current output queue levels. The central unit monitors overall view of congestion and update their random early detection

parameters according to the congestion notification of the control unit [8]. Strategic (RED) Random early detection method is proposed in [4] which use queue parameter. By monitoring the queue parameter according to current queue condition and reduces the queuing delay and increase the throughput. Delay time can be reduced substantially if network length is more and sender and receiver are at sufficient distance and increase the throughput. Exiting RED uses a mechanism early detection of Packet drop without waiting to queue overflow, this mechanism inform the sender to reduce the packet transmission rate and also inform the receiver to not to send excessive acknowledgement packets. In strategic RED parameters are varied according to current queue availability which send excessive acknowledgment and solve above problem.

In [2] this paper an ECN-based congestion control algorithm called access point congestion control (APCC) is presented. In this algorithm using both wireless channel load and buffer length as compound congestion indicator. It provides more stable and higher efficient TCP congestion control and achieve time fairness and higher total network goodput. [5] Also presented an ECN scheme with new response strategy that is more aggressive in the short term, but preserve TCP behavior in long term without modifying the router marking rate. An effort [7] is tried to present distributed ECN-based congestion control protocol to which we refer as Double-Packet Congestion Control Protocol (DPCP). This scheme provide more accurate feedback compare to variable structure congestion control protocol (VCP). By extracting the router information into series of packets .It is capable to notify three level of congestion in two ECN bits. In [9] explicit congestion notification accurately improves the efficiency of TCP without harming its performance. The Probabilistic Congestion Notification (PCN) is designed in [6] to help determine the exact level of congestion at each intermediate queue. This scheme use single bit explicit congestion notification (ECN) in the IP header. The source to estimate the exact level of congestion at each intermediate queue. By knowing this, the source mitigate its sending rate or choose an alternative roots. The estimation mechanism makes use of time series analysis both to improve the quality of the congestion estimation and to predict, ahead of time, the congestion level which subsequent packets will encounter [6].

[3] Have presented a new Code point mechanism that replace the current ECN mechanism and reuse the assigned bit in the TCP header. In order to observe the congestion situation and provide a better accuracy against losses of packets. This code point scheme performs well in internet link scenario and provides more accurate ECN feedback. By reusing the ECN TCP header bits without allocating further option space. In this paper [1], an innovative method variable structure congestion control with bandwidth estimation (VCP-BE) is presented. That uses available end-to-end bandwidth estimation to provide more accurate congestion feedback. With estimated available bandwidth and ECN feedback window adjustment, the algorithm convergence fast and to improve the network performance. But this algorithm suffers from overestimation of bandwidth and difficult computation in routers.

III. PROPOSED APPROACH

In this work enhanced explicit congestion notification technique is used. The proposed enhanced version employs extra bit to provide additional traffic information at the senders end. This additional information allows more flexible handling of congestion. The additional traffic information carried in the extra ECN bit, allows the node to determine more suitable adjustments to the congestion window size, with added options to RED queue to handle congestion more effectively. Conventionally, size of sliding window is decided upon successful delivery acknowledgement of packets and it gradually increases and decreases accordingly. In earlier works, RED queue was administered by two values of minimum and maximum threshold value, but here in this work RED queue is managed by introducing two more threshold value which are depicted by *slow decrement* and *fast decrement* along with similar vice versa pattern, due to highly dynamic nature of MANETS.

The network state is based on the network load factor ρ_l which can be estimated very easily at each router and is associated with following illustrated ECN bits in the IP header of each data packet. In our approach, for every time interval t_p a router observes its input traffic and calculates the load factor ρ_l as-

$$\rho_l = \frac{\lambda + k.q}{\gamma.C.t_p}$$

Where λ and q represent the amount of input traffic and the persistent queue length during the period t_p , k controls how quickly the persistent queue length can be reduced to zero. γ Denotes the target utilization and C indicates the channel capacity.

To adjust size of sliding window, an IMP will maintain a RED queue with four threshold level i.e. min, slow min, slow max, max which when get overburdened and crosses the max threshold level, set both ECN bits which signifies to reduce the window size significantly (let's say, four times) and if RED queue signals slow max threshold level, it reduces the size of window but less than aforesaid case (let's say two times). Below illustrated table expounds the behavior of RED queue which leads to change in ECN bits and adjustment in window size to handle congestion.

Following the aforesaid approach, as and when congestion experienced, the sliding window size is accordingly handled by the control bits sent by intermediate router to the source node. The novelty of the proposed approach lies in using additional traffic information, carried in extra ECN bit, to adjust window size more appropriately, instead of drastically slashing it down. As we are using the Mobile ad hoc network platform which is highly dynamic in nature and connectivity of any node changes abruptly, which sometimes leads to sudden increase/decrease of window size, but sometimes the same is needed to lesser extent, and such situation is handled in a better way in the proposed approach.

TABLE 4.1. BEHAVIOR OF RED QUEUE AND ECN BITS

| RED queue threshold level | ECN bits status | Action to be taken at senders side |
|---|-----------------|--|
| slow ($0 \leq \rho_i \leq 0.7$) | 00 | Increase window size significantly (4 times) |
| slow min ($0.7 \leq \rho_i \leq 0.9$) | 01 | Increase window size gradually (2 times) |
| slow max ($0.9 \leq \rho_i \leq 1.0$) | 10 | Decrease window size gradually (2 times) |
| max ($\rho_i > 1.0$) | 11 | Decrease window size significantly (4 times) |

IV. SIMULATION AND RESULTS

To evaluate the effectiveness of the proposed approach, aforesaid concept has been implemented over a simulation environment using NS2 simulator. In the simulations, we use the $400 \times 400m^2$ gride and set the data packet size as 512B. The basic setting is an 11Mbps link with a 80ms RTT where the forward and reverse path each has 5FTP flows unless stated otherwise. RED is always used with ECN enabled at the routers. All simulations are run for at least 120s. The effectiveness of the approach is being justified by comparing our approach with drop tail approach in literature work illustrated on the basis of Goodput, which is the application level throughput, i.e. the number of useful information bits delivered by the network to a certain destination per unit of time. The amount of data considered excludes protocol overhead bits as well as retransmitted data packets. This is related to the amount of time from the first bit of the first packet sent (or delivered) until the last bit of the last packet is delivered.

Another metrics is average queue length which is a crucial count to measure congestion. A low value of average queue length a better congestion control mechanism is being used. Lastly, loss rate is being considered for the same which is defined as the successfully delivered packets to the total sent packets.

Following mentioned graphs illustrate the effectiveness of our approach as compare to its counterpart drop tail.

From figure 5.1, it is clearly depicted that our approach named “MyRED” is performing better consistently over increasing number of flows.

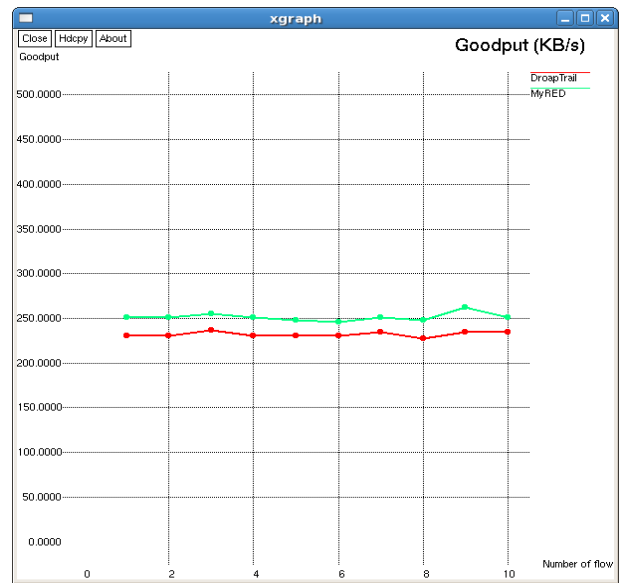


Fig 5.1. Goodput comparison of proposed approach

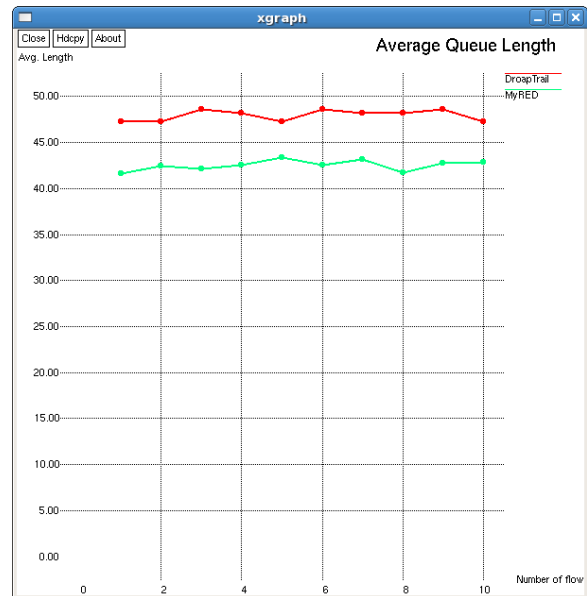


Fig 5.2. Average queue length comparison of proposed approach

Figure 5.2 expounds comparison of average queue length between proposed approach v/s drop tail and here also, our proposed approach MyRED is showing better results than its counterpart. Although it is fluctuating too much but still it is too low when compared to Drop tail approach.

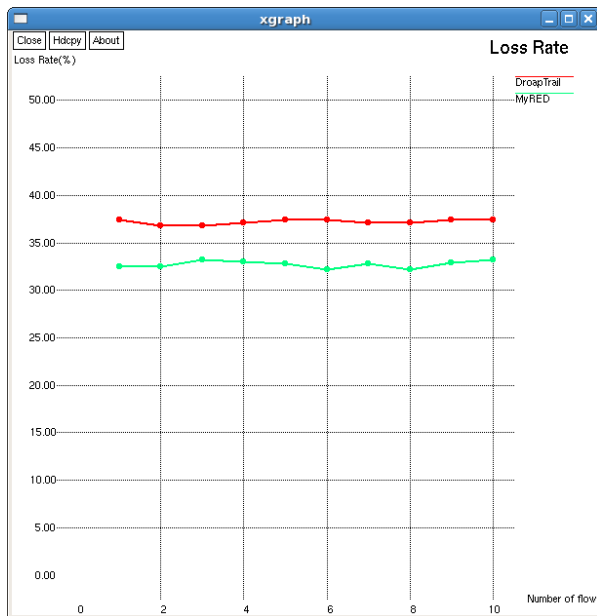


Fig 5.3. Loss rate comparison of proposed approach

Lastly the loss rate observed in simulation setup is also far better (lesser) in our approach in comparison to that of Drop tail approach.

V. CONCLUSIONS

In the aforesaid study and evaluation of approach, we found our approach more effective and suited for congestion

prone networks to handle the same using ECN bits associated with load factor. In all above expounded parameters, MyRED approach is performing better than its variant. The work can be extended and enhanced by implementing it in real hardware to study its strength and flaws in more practical settings.

REFERENCES

- [1] Jianxin Wang, Jie Chen, Shigeng, Weiping Wang, "An explicit congestion control protocol based on bandwidth estimation" 978-1-4244-9268-8/11/\$26.00, IEEE, 2011.
- [2] Jiawei Huang, Jinaxin Wang, "An ECN-Based Congestion Control Algorithm for TCP Enhancement in WLAN" 11th IEEE International Conference on High Performance Computing and Communications, 2009.
- [3] Mirja K"uhlewind, Richard Scheffenegger, "Design and Evaluation of Schemes for More Accurate ECN Feedback" Workshop on Telecommunications 2012.
- [4] Jeetendra Kr Patel, Jigyasu Dubey, "Mobile Ad hoc Network Performance Improvement Using Strategical RED" 978-1-4673-1989-8/12/\$31.00, IEEE, 2012.
- [5] Minseok Kwon and Sonia Fahmy, "TCP Increase/Decrease Behavior with Explicit Congestion Notification" 0-7803-7400-2/02/\$17.00 © 2002 IEEE.
- [6] Mussie Woldeselassie, Richard G. Clegg, Miguel Rio, "Forecasting Full-Path Network Congestion Using One Bit Signalling" 978-1-4244-6404-3/10/\$26.00 ©2010 IEEE.
- [7] Xiaolong Li, Homayoun Yousefi Zadeh, "Distributed ECN-based congestion control" 978-1-4244-3435-0/09/\$25.00, IEEE, 2009.
- [8] Ozen Yelbasi, Emin German, "A New Approach to Estimate RED Parameters Using Global Congestion Notification", International Conference on Network Computing and Information Security, 2011.
- [9] Stefanos Harhalakis, Nikolaos Samaras, Vasileios Vitsas "An experimental study of the efficiency of Explicit Congestion Notification" 978-0-7695-4389-5/11 \$26.00 © 2011 IEEE.

Clustering of Image Data Using K-Means and Fuzzy K-Means

Md. Khalid Imam Rahmani¹

¹Associate Professor, Deptt. of Computer Sc. & Engg.
Echelon Institute of Technology
Faridabad, INDIA.

Naina Pal², Kamiya Arora³

^{2,3}M.Tech. Scholar, Deptt. of Computer Sc. & Engg.
Echelon Institute of Technology
Faridabad, INDIA.

Abstract—Clustering is a major technique used for grouping of numerical and image data in data mining and image processing applications. Clustering makes the job of image retrieval easy by finding the images as similar as given in the query image. The images are grouped together in some given number of clusters. Image data are grouped on the basis of some features such as color, texture, shape etc. contained in the images in the form of pixels. For the purpose of efficiency and better results image data are segmented before applying clustering. The technique used here is K-Means and Fuzzy K-Means which are very time saving and efficient.

Keywords—Clustering; Segmentation; K-Means Clustering; Fuzzy K-Means

I. INTRODUCTION

Clustering is the unsupervised classification of patterns such as observations, data items, or feature vectors into groups named as clusters [1]. Applications of clustering is growing nowadays very rapidly because it saves a lot of time and the results obtained from the clustering algorithm is very suitable for the algorithms in the later stages of the applications. Clustering basically groups the data. The data in every group is similar to each other but quite dissimilar to the data in different groups [5]. So, the data which are grouped together are similar to each other.

Clustering has very wide range of applications in the field of research & development like in medical science, where the symptoms and cures of diseases are grouped into clusters to save time and achieve efficient results [10]. It is applied in image processing, data mining and marketing etc. In information retrieval clustering can enhance the performance of retrieving of information from the Internet considerably. All pages are grouped into clusters and optimal results are achieved.

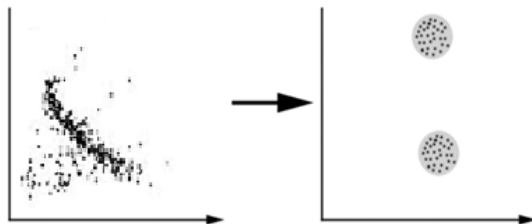


Fig. 1. Grouping of Similar Data Points

Clustering may also be used in marketing scenarios as it can segment the market into many profitable groups including advertising, promotions and follow ups etc [10]. Clustering can also be used in archeology where researchers are trying to discover stone tools, funeral tools etc. to save time in investigation surveys [10].

Image clustering can also be used in order to segment a movie [4]. Clustering is defined as unsupervised learning where user can randomly selects the data points without the help of a supervisor. There are huge applications of clustering as data clustering has proved a very powerful technique in classifying each application into clusters and sub-clusters for easy, quick and efficient results [11].

A brief description of the state of the art of clustering and various forms of clustering are given in section II. K-Means applied on image is described in section III. In section IV, an overview of existing methodologies has been described. Segmentation of images is being described in section V. In section VI, a proposed algorithm has been described. Section VII has been used for the conclusion and future direction of the research work.

II. THE STATE OF THE ART

A. Clustering

Clustering is a method which groups data into clusters, where objects within each clusters have high degree of similarity, but are dissimilar to the objects in other clusters. So, Clustering is a method of grouping data objects into different groups, such that similar data objects belong to the same cluster and dissimilar data objects to different clusters [9]. Clustering involves dividing a set of data points into non-overlapping groups or clusters of points where points in a cluster are “more similar” to one another than the points present in other clusters [2]. Clustering of images is done on the basis of the intra-class similarity. Target or close images can be retrieved a little faster if it is clustered in a right manner [8]. Data points in each cluster are calculated with a data points in the cluster, similar data points are brought in one cluster. So, each data points exhibits same characteristics present in one cluster.

So, a good clustering method would exhibit high similarity in a single cluster and a very less similarity with other clusters.

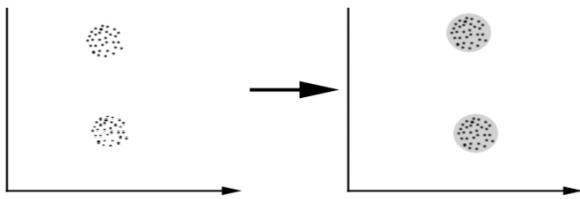


Fig. 2. Clustering of Data Points

In 1997, Haung brought the concept of k-modes which was the extension made on k-means algorithm. K-modes algorithm was introduced to cluster the numeric objects.

In 1999, Guha et al proposed a clustering algorithm of number of links between tuples. These links were used to captures the records and used to describe that which records are similar with each other. It gave satisfactory results.

In 2005, FUN and Chen presented KPSO clustering algorithm which merges some ideas of k-means and PSO. This was proposed to automatically detect the centroids of the cluster of geometric structure data sets.

In 2006, Csorba and Vajk introduced a document clustering method in which there was no need to assign all the documents to the cluster, only relevant documents were being assigned to the cluster. So, it leads to the cleaner results.

In 2007, Jing et al introduced a new k-means technique for the clustering of high dimensional data. So, different topic documents are placed with the different keywords.

B. K-Means Clustering

K-Means algorithm is the most popular partitioning based clustering technique. It is an unsupervised algorithm which is used in clustering. It chooses the centroid smartly and it compares centroid with the data points based on their intensity and characteristics and finds the distance, the data points which are similar to the centroid are assigned to the cluster having the centroid. New 'k' centroids are calculated and thus k-clusters are formed by finding out the data points nearest to the clusters.

Steps of the K-Means [10] algorithm can be outlined as mentioned below:

1. Choose k number of points randomly and make them initial centroids.
2. Select a data point from the collection, compare it with each centroid and if the data point is found to be similar with the centroid then assign it into the cluster of that centroid.
3. When each data point has been assigned to one of the clusters, re-calculate the value of the centroids for each k number of clusters.
4. Repeat steps 2 to 3 until no data point moves from its previous cluster to some other cluster (termination criterion has been satisfied).

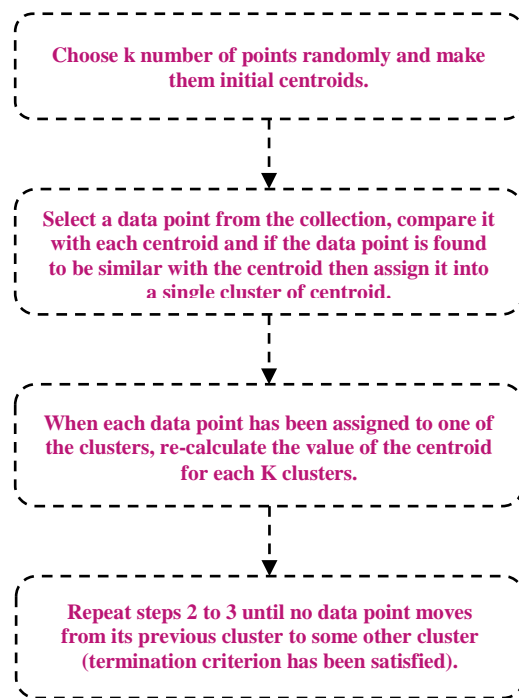


Fig. 3. K-Means Clustering Algorithm

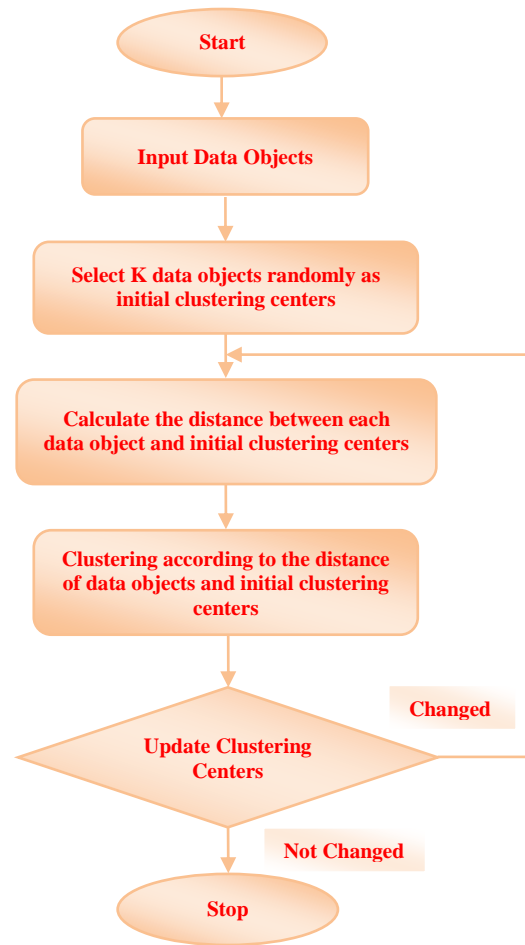


Fig. 4. K-Means Flow Chart

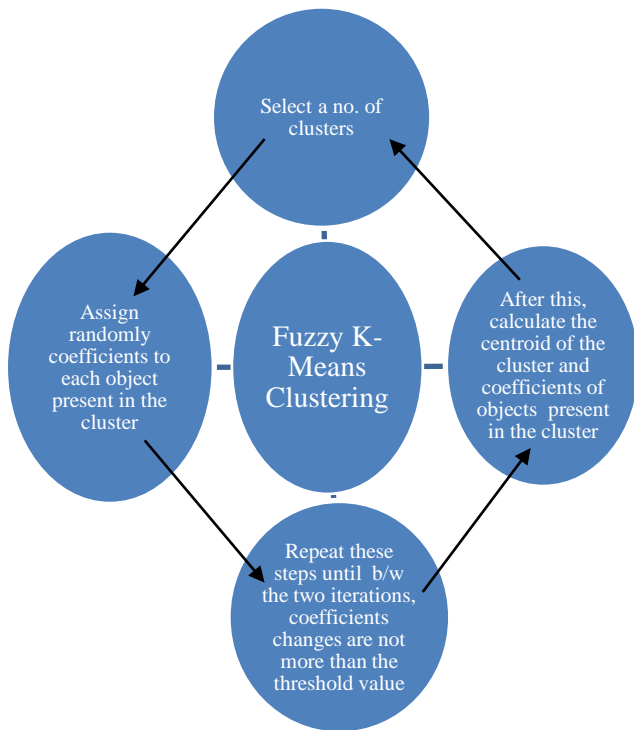


Fig. 5. Fuzzy K-Means Flow Chart

C. Fuzzy K-Means Clustering

Each object in the fuzzy clustering has some degree of belongingness to the cluster. So, the objects that are present on the edge of the cluster are different from the objects that are present in the centroid i.e. objects on edge have lesser degree than the objects in the center. Any object p has assigned a set of coefficients that are present in the k -th cluster $w_{k-1}(p)$. In the fuzzy c -means [1], the centroid of a cluster is the mean of all objects present in the cluster, measured by their degree of belonging of points to the cluster:

$$C_{k-1} = \frac{\sum_p w_{k-1}(p)^m p}{\sum_p w_{k-1}(p)^m}$$

III. K-MEANS APPLIED ON IMAGE DATA

Let us elaborate this by taking an example, DSR (Dynamic Spatial Reconstructor) scans left atrium and also there is a left ventricle which includes aorta and chamber [1]. Although there are valves separating the left ventricle chamber from left atria chamber and aorta still visibility is diminished because of the limitation of DSR. This disadvantage of DSR in medical image system is removed by K-means as K-means calculates the intensity of every pixel and then makes clusters. So, K-means proposed a cluster corresponding to the brightest regions would represent the left ventricle chamber and left ventricle chamber visibility becomes bright [1].

The fuzzy k-means algorithm is very similar to the k-means algorithm as depicted in figure 5 in the previous column.

IV. OVERVIEW OF METHODOLOGIES USED

Hartigan (1975) defines clustering as the group of similar data objects. The goal of clustering is to partition the data sets into several groups based on its similarities or dissimilarities i.e. entities that belong to a single group are considered to be similar to each other.

There has been much advancement done in clustering using k-means. Some of the advancements are given below in various fields of clustering:

Hierarchical clustering was introduced which is the one which makes hierarchies of cluster and these hierarchies of clusters are made to form a tree of clusters are known as dendrograms. There are two types of hierarchical clustering one is agglomerative method and the second one is divisive method. In agglomerative method, each object makes a cluster and the two most similar clusters are merged with each other and they merge iteratively up to a single cluster with the objects has been formed.

Agglomerative is based on the inter cluster similarity whereas in Divisive method, cluster is selected and splits up into many smaller clusters recursively until some termination criterion has been obtained.

Partitioning clustering introduced which splits objects into many subsets. It uses some greedy heuristics. Partitioning clustering has a drawback that many output clusters are being formulated. Berkhin (2006) describes it as a major advantage of partitioning clustering, the fact that iterative optimization may gradually improve clusters. This would result in high quality clusters. This is unlike hierarchical clustering, as algorithms in which class do not feature re-visits to clusters.

K-means clustering introduced K-Means is also known as straight K-means originated independently in the works of MacQueen (1967) and Ball and Hall (1967). Clustering came in the research since the 1960s. Factor analysis was the first related work took place by scholars (Holzinger, 1941), numerical taxonomy (Sneath and Sokal, 1973), and unsupervised learning in pattern recognition (Duda and Hart, 1973). Nowadays, clustering is used in many fields. In bioinformatics, Clustering is widely used in microarray data analysis (Allison et al., 2006) also bioinformaticians mostly use clustering because of this reason researchers have compared clustering algorithms within the field (Yeung et al., 2001), including the problem of K-means initialization. One who is working within the field of computer vision is also become the keen user of K-means, an example of the use of K-Means in this context would be to cluster the entities in an image (Szeliski, 2010) based on each pixel's features: normally their color and position. K-Means is the most popular clustering algorithm, which generates the non-overlapping clusters. It is more efficient than the hierarchical algorithms (Manning et al., 2008). K-means has been used to solve much number of problems since the 1960s. Each cluster has a centroid which is used to represent the general features of the cluster, it basically chooses any random centroid and assigns data points to the centroid by comparing distance of the data points with the centroid, the data points which has least distance with the centroid are made to form one cluster. It

computes k-centroids by using this process and changes k-centroids values iteratively until some termination criterion has been obtained.

V. SEGMENTATION OF IMAGES

Image segmentation has attracted considerable attention for the last few years, due to the advances in multidimensional image acquisition techniques [3].

Image segmentation is used to represent some characteristics, features from images. Segmentation operated on the images segments the images, extracts some of its important features and matches these features and matches these features with the pixels of the images. Efforts have been made to segment an entire volume (rather than merging a set of segmented slices) using supervised pattern recognition techniques or unsupervised fuzzy clustering [6]. The similar one makes one cluster and the similar cluster is dissimilar from another clusters.

VI. PROPOSED WORK

The work which has to be done combines some ideas of image segmentation into content based image classification. In this work, the concept of image segmentation for medical images using techniques of clustering is being proposed. Retrieval of images based on segmentation and clustering of images gives better results. Here, in this it is being proposed to focus on some feature selection and moreover on classification of medical image data which is based on some of the feature selection algorithm.

VII. CONCLUSION

Fuzzy k-means is better than k-means by many factors like first, it give better results when compared with k-means algorithm by increasing the fuzzy factor. Secondly, Fuzzy K-means takes lesser time to cluster the images than K-means. Thirdly, K-means is considered to be a hard clustering and in hard clustering, after some iteration most of the centers are converged to their final positions and the majority of data points has only few candidates to be selected as their closest centers where as Fuzzy K-means is known as soft clustering in which the data points which are present in the fuzzy K-means can belong to more than one cluster with having certain probability. Moreover the distance measured by the k-means is considered to be a distortion measured and the distance measured has been extended to the fuzzy K-means.

REFERENCES

- [1] A.K. Jain, M.N. Murty and P.J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, Vol. 31(3), September 1999, pp. 264-323.
- [2] Vance Faber, "Clustering and the Continuous k-Means Algorithm", Los Alamos Science, Vol. 22, 1994, pp. 138-144.

- [3] Chang Wen Chen, Jiebo Luo and Kevin J. Parker, "Image Segmentation via Adaptive K-Mean Clustering and knowledge-Based Morphological Operations with Biomedical Applications", IEEE Transactions on Image Processing, ISSN: 1057-7149, Vol. 7(12), December 1998, pp. 1673-1683.
- [4] Yevgeny Seldin Sonia Starik Michael Werman, "Unsupervised Clustering of Images Using their Joint Segmentation", pp. 1-24.
- [5] Madhuri A. Tayal, M.M. Raghuvanshi, "Review on Various Clustering Methods for the Image Data", Journal of Emerging Trends in Computing and Information Sciences, ISSN: 2079-8407, Vol. 2, 2011, pp. 34-38.
- [6] Clark MC, Hall LO, Goldgof DB, Clarke LP, Velthuizen RP, Silbiger MS, "MRI Segmentation Using Fuzzy Clustering Techniques", IEEE Engg Medicine and Biology, ISSN: 0739-5175, Vol. 13(5), Dec 1994, pp. 730-742.
- [7] Pham DL, Prince JL, "Adaptive Fuzzy Segmentation of Magnetic Resonance Image", Vol. 18(9), Sep 1999, pp. 737-752.
- [8] A.Kannan, Dr.V.Mohan, Dr.N.Anbazhagan, "Image Clustering and Retrieval using Image Mining Techniques", 2010 IEEE International Conference on Computational Intelligence and Computing Research, 2010.
- [9] Dr. Sanjay Silakari, Dr. Mahesh Motwani, Manish Maheshwari, "Color Image Clustering using Block Truncation Algorithm", International Journal of Computer Science Issues, ISSN: 1694-0784, Vol. 4(2), 2009, pp. 31-35.
- [10] P. Bradley, and U. Fayyad, "Refining Initial Points for K-Means Clustering," In Proceeding of 15th International Conference on Machine Learning, Jan 1998, pp. 91-99.
- [11] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques", 2006.

AUTHORS' BIBLIOGRAPHY



Md. Khalid Imam Rahmani is an Associate Professor in the Department of Computer Science & Engg. of a very reputed NBA accredited Engineering College, Echelon Institute of Technology, Faridabad, India. He is having about 17 years of teaching, industry and administrative experience. He has done B.Sc. Engg. in Computer Engineering from A.M.U., Aligarh, M.Tech. in Computer Engineering from M.D.U., Rohtak and is pursuing Ph.D. in Digital Image Retrieval Algorithms. Digital Image Processing, Innovative Programming techniques, Mobile Computing, Algorithms Design and Internet & Web Technologies are his research areas.



Naina Pal has earned her M.Tech. degree in Computer Science & Engg. from Echelon Institute of Technology under Maharshi Dayanand University, Rohtak. Her research interests include Image Processing, Classification of data using Clustering and Data mining.



Kamiya Arora has earned her M.Tech. degree in Computer Science & Engg. from Echelon Institute of Technology under Maharshi Dayanand University, Rohtak. Her research interests include Image Processing, Steganography, Data mining and Cryptography.

Identifying and Extracting Named Entities from Wikipedia Database Using Entity Infoboxes

Muhidin Mohamed

School of Electrical, Electronic and Computer Eng.
University of Birmingham,
Edgbaston, Birmingham, United Kingdom

Mourad Oussalah

School of Electrical, Electronic and Computer Eng.
University of Birmingham
Edgbaston, Birmingham, United Kingdom

Abstract—An approach for named entity classification based on Wikipedia article infoboxes is described in this paper. It identifies the three fundamental named entity types, namely; Person, Location and Organization. An entity classification is accomplished by matching entity attributes extracted from the relevant entity article infobox against core entity attributes built from Wikipedia Infobox Templates. Experimental results showed that the classifier can achieve a high accuracy and F-measure scores of 97%. Based on this approach, a database of around 1.6 million 3-typed named entities is created from 20140203 Wikipedia dump. Experiments on CoNLL2003 shared task named entity recognition (NER) dataset disclosed the system's outstanding performance in comparison to three different state-of-the-art systems.

Keywords—named entity identification; Wikipedia infobox; infobox templates; Named Entity Classification (NEC);

I. INTRODUCTION

The word named entity (NE) as used today in text mining and Natural Language Processing (NLP) was introduced in the Sixth Message Understanding Conference [1]. It represents a major part of all textual data covering proper names of persons, locations, organisations and corporate entities e.g, University of Birmingham, UK, Mount Everest, Mogadishu, David Beckham among others. Besides, Named entity classification (NEC) is the process of categorizing named entities to their corresponding classes (e.g. Person, Location, Organization). This is usually a supplementary step to the wider area of named entity recognition (NER). Although, NEs represent core components in natural language texts, they are still poorly covered in the state of the art language dictionaries. This might be due either to their ever-changing nature and dynamicity in which some named entities disappear while new ones emerge on regular basis, or to the fact that many NEs might be genuinely classified to more than one class, where one may encounter, for instance, several place names who are also person names, and/or corporate names. For example, if you search some of world's largest corporations such as Microsoft and Apple you may hardly find them in the state of the art knowledge networks such as WordNet. An improvement of named entity coverage are now being made in lexical semantic networks such as ConceptNet 5 [2]. More importantly, constantly updated live online repositories like Wikipedia [3] and Open Directory Project [4] do possess high named entity coverage than the aforementioned resources holding almost all object names. Therefore, in order to automatically handle NER or NEC tasks, the use of such repositories is inevitable.

Challenges hindering an accurate NEC is not limited to their low coverage in the well-established language resources, but also include the ambiguity pervading the meaning of these entities [5], and entity linking [6], which have been subjected to intensive studies in recent years. This study is rather focused on improving NEC through addressing the coverage problem. To this end, current work advocates the use of Wikipedia utility for entity classification.

Strictly speaking, with the emergence of diverse natural language processing tools and the increasing need for automated text analysis, an important research has been conducted for the purpose of named entity classification in the past few years. In [7], authors used a bootstrapping method based on Wikipedia category to classify named entities containing Heidelberg Named Entity Resource (HeiNER) [8]. Nevertheless such classification might be undermined by the inconsistency of placing contributed articles by the authors in the most appropriate category. In a closely related study, Tkachenko et al. [9] carried out a fine grained classification for Wikipedia named entities. Though, their method correlates this study, they extracted many features for the classification including first paragraph of the article text, categories, template names, and other structured content tokens. This will demand a huge processing time when classifying large datasets. The closest work to ours is explored in [10] where researchers used structured information from infoboxes and category trees for the classification task. Despite this relatedness, their work differs from this study in terms of the overall classification methodology as well as the employed dataset where Portuguese Wikipedia was used in [8].

Finally, one shall also mention some seminal works on Wikipedia entity classification built on machine learning algorithms. Dakka et al. [11] used bag-of-words of Wikipedia articles with support vector machine (SVM) algorithm achieving a high F-score of (90%). Watanabe et al. [12] employed Conditional Random Fields to classify Japanese Wikipedia articles while Bhole et al. [13] combined heuristics with linear SVM for the same purpose. But the main drawback of machine learning related approaches lies in the requirement of a manually annotated training data, which is rather costly and complex task.

The main contribution of this paper consists in designing and testing a new simple named entity classification algorithm that only makes use of some structured information available in Wikipedia articles. Especially, unlike the aforementioned methods, the proposed NEC approach relies on the content

information of a single structured table, the infobox, but achieves a high score of accuracy and F-measure. The classification algorithm put forward in this study matches a predefined core entity attributes built from Wikipedia Infobox Templates (WIT) and entity specific attributes extracted from the related named entity Wikipedia article.

The rest of the paper is structured as follows. Section 2 covers Wikipedia structure and its containment of named entities. Section 3 copes with the proposed named entity classification approach using Wikipedia. Section 4 details the system experiments, highlighting the utilized dataset, results, and comparison with relevant state of the art systems. Finally, conclusions are drawn in Section 5.

II. WIKIPEDIA

A. Overview

Wikipedia is a freely available encyclopaedia with a collective intelligence contributed by the entire world community [14]. Since its foundation in 2001, the site has grown in both popularity and size. At the time of this study's experiment (April 2014), Wikipedia contains over 32 million articles [15] in 260 languages [16] where its English version has more than 4.5 million articles¹. Its open collaborative contribution to the public arguably makes it the world's largest information repository.

Wikipedia contains 30 namespaces of which 14 are subject namespaces and two are virtual namespaces. Besides, each namespace has a corresponding talk namespace². A namespace is a criterion often employed for classifying Wikipedia pages, using MediaWiki Software, as indicated in the page titles. Structurally, Wikipedia is organized in the form of interlinked pages. Depending on their information content, Wikipedia pages are loosely categorized as Named Entity Pages, Concept Pages, Category Pages, Meta Pages [8].

In recent years, there has been a growing research interest among the NLP and IR research communities for the use of this encyclopaedia as semantic lexical resources for tasks such as word semantic relatedness [17], word disambiguation [18], text classification [19], ontology construction [20], named entity recognition/classification [21], among others.

B. Named Entities in Wikipedia

Research has found that around 74% of Wikipedia pages describe about named entities [22], a clear indication of Wikipedia's high coverage for named entities. Each Wikipedia article associated with a named entity is identified with its name. Most Wikipedia articles on named entities offer useful unique properties starting with a brief informational text that describes the entity, followed by a list of subtitles which provide further information specific to that entity. For example, one may find information related to main activities, demography, and environment for Location named entities; education, career, personal life and so on for Person named entities. Relating concepts to that named entity are linked to the entity article by outgoing hyperlinks. Moreover, a semi-

structured table, called infobox, summarizing essential attributes for that entity lives in the top right hand of each article [23]. It is the core attributes of the article infobox that this study stands on for the classification of named entities without any other prior knowledge. The snapshot in Figure 1 illustrates the Wikipedia article infobox related to "Google", which corresponds to a named entity of type Organization (<http://en.wikipedia.org/wiki/>). The table summarizes very important unique properties of the entity in the form of attribute-value pairs. Consequently such tables are extracted, stored and analysed for the purpose of NE classification.



Fig. 1. Google Wikipedia article infobox³

III. THE CLASSIFIER

Using predefined core attributes extracted from Wikipedia Infobox Templates, a semi-supervised binary algorithm is developed. Being the main classifier, it predicts whether a particular named entity belongs to a given type. In other words, the classifier is designed to match named entities against these set of core class attributes (cf. Section A) and consequently identify these entities based on the outcomes of the matching process. The classification is achieved according to the following definition.

¹ http://en.wikipedia.org/wiki/Main_Page

² <http://en.wikipedia.org/wiki/Wikipedia:Namespace>

³ <http://en.wikipedia.org/wiki/Google>

Definition: Let ne be a named entity in Wikipedia (**WP**) belonging to any of the three types, Person (**P**), Location (**L**) and Organization (**O**). If **XITA** denotes infobox template attributes⁴ of type **X** and **IA**(ne) is the infobox attributes extracted from WP article associated with ne , then the classifier identifies ne type according to quantification (1).

$$T_{ne} = \begin{cases} P & \text{if } ne \in WP \ \& \ IA(ne) == PITA \\ L & \text{if } ne \in WP \ \& \ IA(ne) == LITA \\ O & \text{if } ne \in WP \ \& \ IA(ne) == OITA \end{cases} \quad (1)$$

Where T_{ne} stands for the type of named entity ne as identified by the classifier, while the operator “==” corresponds to array matching.

A. Defining Core Attributes

MeidaWiki team has developed infobox templates designed to guide contributing authors. The infobox templates contain the attribute labels to be filled by the authors with values when writing their Wikipedia articles on named entities. These attributes describe properties particular to each named entity type. For example, all location-based named entities should bear **coordinate** information. Similarly, infobox attributes for Person named entities include **birth date** and **place**. Table 1 lists a selected sample of these attributes for demonstration purpose. Essential attributes to each class, usually identified through manual investigation, are referred **Core Attributes**. The latter are used in the experiments to identify Wikipedia articles corresponding to named entities through matching the core attributes with the attributes extracted from entity infoboxes. Experimented core attributes are designated with stars in Table 1.

TABLE I. CORE ATTRIBUTES EXTRACTED FROM INFOBOX TEMPLATES

| Person | Organization | Location |
|--------------|------------------------|--------------|
| Birth_date* | Ceo, Founded* | Coordinates* |
| Birth_place* | Headquarters* | Population* |
| Spouce | Service_area* | Area* |
| Children | Industry , Profit* | Region |
| Relatives | Traded_as, revenue* | Country* |
| Occupation | Num_staff*, | timezone |
| Nationality | Num_employee* | iso_code |
| Parents | Established* | area_code |
| Education | Founder/chancellor* | settlement |
| Salary | {Post under}graduates* | Leader_name |
| partner | {operating net}income* | Leader_name |

B. Accessing Wikipedia Database

To use Wikipedia as an external knowledge repository for named entity classification, a mechanism for accessing its database should be in place. Designed system’s access to the encyclopaedia is summarized in Figure 2. Primarily there are two methods for accomplishing such data access; namely, either querying through web interface, or accessing a downloaded local Wikipedia dump.

For this study, query access method is used for the system evaluation. However, for the actual named entity extraction, a local access is made to a downloaded Wikipedia xml dump of

February 2014. In implementing the query access method, this study partially adapts the Wikipedia Automated Interface [24] while the local access to the Wikipedia Dump is built on a MediaWiki dump Files Processing Tool [25]. The preference of query access over the local access for the evaluation is tied to the unsuitability of the dump files for random access as the dumps are primarily designed for sequential access.

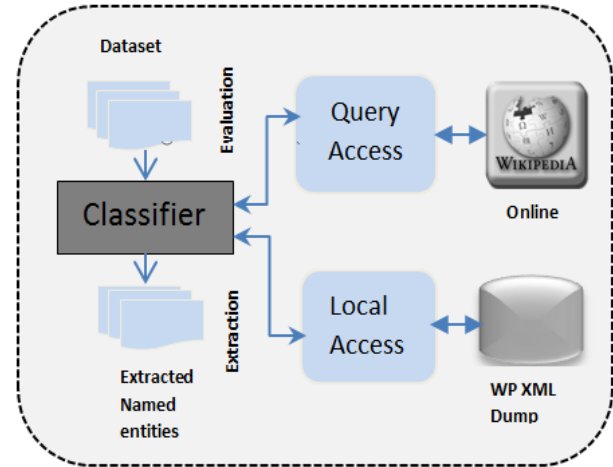


Fig. 2. Classifier’s Access Mechanisms to Wikipedia

IV. EXPERIMENTAL SETUP

The proposed classifier system is implemented with Perl scripts in Linux environment. Entity core attributes derived from Wikipedia Infobox Templates represent the heart of the developed classification method. An illustration of the implementation scheme is given in Figure 4 (cf. the algorithm in Fig. 3). Each named entity has to go through three processing stages before it gets classified to its type. In stage 1, the Wikipedia article associated with that entity is retrieved while the extraction of its article’s infobox forms stage two. At this stage, the scope of the processing text has been narrowed to the infobox. This semi-structured table is further parsed in stage 3 where tuples of attribute label-values are built from the infobox obtained in stage 2. Having organized the tuples in Perl Hashes, the matching process is now performed against the core attributes and the correct decision is made. The same process is repeated for every named entity to be identified. Figure 3 and Figure 4 better summarize the logical flow of the discussed classification methodology, in terms of pseudo-code and block diagram representation.

Algorithm1 WP Aided NE Classification

- 1 ED ← NE Evaluation Dataset
- 2 AV ← Infobox template Attributes
- 3 C ← {}
- 4 For all ($ne_i \in ED$) do
- 5 If $ne_i \in WPDB$ then
- 6 $A_{ne_i} \leftarrow$ RetrieveArticle(ne_i)
- 7 $I_{ne_i} \leftarrow$ ExtractInfobox(A_{ne_i})
- 8 For each $v_j \in AV$
- 9 If $v_j \sim I_{ne_i}$ then

⁴ These are the core attributes used for matching

```
10     cne ← ne, #type(vi)  
11     Last;  
12     Endif  
13     Endfor  
14 endif  
15 C ← C ∪ { cne }  
  
16 endfor  
17 return C
```

Fig. 3. Perl-styled Pseudocode algorithm for Wikipedia infobox-based named entity classification.

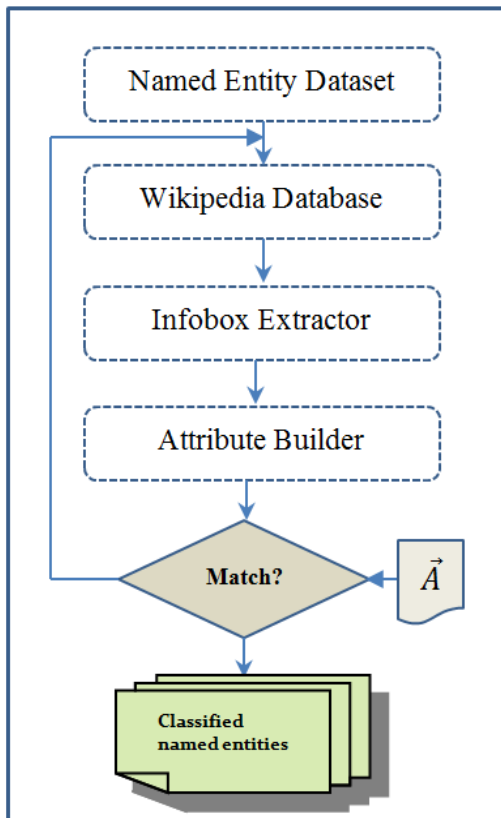


Fig. 4. Named entity Classifier Flowchart

A. Dataset

Experiments were conducted on two datasets. The first test-data comprises 3600 named entities with different proportions of the three considered entity types (PER, LOC, ORG), and was created from two data sources; namely, Forbes and GeoWordNet. Especially, all organization and person names were an excerpt of Forbes400 and Forbes2000 lists for richest American businessmen and world's leading public companies respectively⁵. On the other hand, Location named entities were sourced from GeoWordNet database. The second test-data uses CoNLL-2003 shared task named entity data⁶. The latter dataset, a standard publicly available dataset, has been selected

⁵ <http://www.forbes.com/lists/>

⁶ <http://www.cnts.ua.ac.be/conll2003/ner/>

for proper evaluation and comparison with state of the art techniques for Wikipedia NEC. Checking the coverage and the availability of all names with their surface forms in Wikipedia has been performed over all datasets prior to the experiments.

B. Results and Discussion

The system tests were made in two rounds. In the first round the test dataset is divided into 4 smaller parts containing 100, 500, 1000, 2000 NEs all with different proportions of their types. This splitting has been performed for at least two reasons. First, this helps to securitize the data size effect on the observed parameters. Second, it reduces Wikipedia server's overhead with large data since all the testing and evaluation experiments used Query-based access to the online version of the encyclopaedia.

There are four possible outcomes that can result from the binary predictive classifier. In the first case, an entity that belongs to a type x might be classified as being of class x , referred to as True Positive (TP). Secondly, A False Negative (FN) occurs when a named entity of type x is incorrectly identified as not falling in that type. Thirdly, there happens a case where a named entity does not belong to class x , but classified as type x ; a situation known as False Positive (FP). Lastly, when a non-member named entity of type x is correctly predicted as not falling in class x , it is referred to as True Negative (TN). Metrics for evaluating the classifier's performance will be based on the above mentioned outcomes.

Results of round 1 experiments are reported in Table 2, where the accuracy level is determined as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

TABLE II. RESULTS: ACCURACY WITH VARYING DATA SIZES

| Dataset Size | Percentage Accuracy | | |
|--------------|---------------------|----------|--------------|
| | Person | Location | Organization |
| 100 | 96% | 99% | 97% |
| 500 | 91.6% | 95.4% | 94% |
| 1000 | 93.8% | 94.2% | 94.3% |
| 2000 | 95.5% | 93.9% | 97.25% |

The trend of the scores shown in Table 2 indicates that varying data sizes has little effect on the accuracy for the Person and Organization entity types. However a slight declination is observable in the case of Location names. Overall, round 1 experiments on test-data reveal that the classifier can achieve an average accuracy above 93% irrespective of the data size.

An examination of the misclassified proportion of the test data showed, that a number of factors contribute to the classifier's failure to identify some named entities. The most prominent factors were found to be the ambiguity of named entities and the absence of infoboxes from Wikipedia articles. Although there are machine learning like solutions to the ambiguity issue, using disambiguation like method, little can be done in the case of absence of infobox information. Possibly, the only sensible way to handle this matter is removing the underlying case (s) from the evaluation dataset in the validation and state of the art comparison stages.

TABLE III. ERROR CAUSING FACTORS AND SYSTEM MISIDENTIFICATIONS

| Type | Ambiguity | No Infobox | Others |
|--------------|-----------|------------|--------|
| Person | 46.7% | 50% | 3.3% |
| Location | 11.1% | 11.1% | 77.8% |
| Organization | 70.1% | 20.8% | 7.1 |

Disambiguation is the process of normalizing named entities that have multiple surface forms and identifying their referents. For instance, Birmingham may refer to the largest city in Alabama USA, or the second largest city in the United Kingdom. Error analysis related to system’s misclassification highlighting factors leading to these errors is presented in Table 3. Through the error analysis, it is found that ambiguity in Organization and Person names extremely undermines the system performance. This is perhaps due to the use of abbreviations for larger organization names and the presence of common cultural names e.g John, Mohamed, shared by thousands of people in Wikipedia database. Disambiguating named entities in Wikipedia has been studied [5] and is still an active research problem.

Results of Table 3 have also shown the existence of a high proportion of Wikipedia named entities that lacked infoboxes information. Experimental results disclosed that **50%** of the unclassified Person entity articles are without infoboxes in Wikipedia. The figure is slightly lower for the other two considered entity types. As the system relies on information in the infobox, the absence of the infobox from any entity article makes the system unable to identify related named entity. Because of its importance, [26] proposed an author assistant tool for automatic suggestion of infoboxes for contributing authors.

In Table 3, the column designated by **Others** combines other factors including redirected pages, and technical difficulty of extracting the infobox due to the structure of some Wikipedia articles that lack regular patterns. Sometimes the availability of an infobox in an article does not guarantee the presence of the core attributes. The fact that some Wikipedia article infoboxes does not contain the core attributes such as coordinates made this factor to be the misclassification culprit for the largest percentage (77.8%) of unclassified Location named entities. This again precluded the classification of these entities on the basis of their core attributes.

Following the error analysis and prior to the second round of evaluative experiments, Wikipedia assisted disambiguation is used to exclude all ambiguous names. Similarly, all named entities whose Wikipedia articles lack infobox tables have been iteratively removed from the evaluation dataset.

In the second round, experiments were conducted using named entities constructed from CoNLL-2003 shared task data for named entity recognition to observe three of the traditional information retrieval metrics namely; precision, recall, and F-measure. Precision is the proportion of classified named entities that belong to the target type. It is defined by the relationship in expression 3.

$$P = \frac{TP}{TP + FP} \tag{3}$$

Likewise, recall (exp. 4) measures the proportion of named entities of a given type which has been correctly classified.

$$R = \frac{TP}{TP + FN} \tag{4}$$

Due to the trade-off between precision and recall, an F-measure has been developed as proper measure that combines the effect of the metrics as formulated in equation 5.

$$F = \frac{2RP}{P + R} \tag{5}$$

The overall classifier results in terms of these three metrics are summarized in Table 4. The F-measure scores of locations and organizations indicate that the selected core attributes represent good classification criteria for identifying Wikipedia entities. Again, this study’s results confirmed that these attributes are mainly added by article contributors when authoring Wikipedia articles through adapting infobox templates. Person names achieved the highest F-score as ambiguity of these has been accounted for.

TABLE IV. OVERALL CLASSIFIER RESULTS

| Type | Precision | Recall | F-score |
|--------------|-----------|--------|---------|
| Person | 1 | 0.98 | 0.99 |
| Location | 0.99 | 0.95 | 0.97 |
| Organization | 0.94 | 0.97 | 0.96 |

C. State-of-the-Art Comparison

Comparing the study’s infobox based matching approach with related state of the schemes for named entity classification and extraction is not trivial. Major discrepancies arise from the peculiarity of each approach in terms of the Wikipedia features (article text, links, categories, infoboxes) used for the entity identification. In addition, there might be significant differences in the evaluation data and Wikipedia language in the event of language dependent schemes. Nevertheless, a rough approximate comparison of the system with three baselines is provided in Table 5. The criteria for choosing these baselines are their closeness to the system in terms of their use of infobox information and related features.

TABLE V. COMPARING F-SCORES WITH BASELINE SYSTEMS

| System | PER | LOC | ORG |
|-------------|------|------|------|
| Bhole [13] | 72.7 | 70.5 | 41.6 |
| Gamallo[10] | 88 | 63 | 73 |
| Tardif [27] | 95 | 99 | 93 |
| This system | 99 | 97 | 96 |

Table 5 compares the outcomes of the overall classification system in terms of F-score for each type of named entity to three state of art classification approaches (baselines). The baselines use infobox data as one of their classification features; whereas this system is entirely built on infobox attribute matching. Despite that, it is evident that it outperforms all baselines except [27] where a high F-score is reported for location based named entities. However, there is still a room for improvement to extend the work in identifying *Miscellaneous* named entities and further subcategorizing the

main entity types to subcategories which have been considered by many state of the art systems.

D. NE Extraction from Wikipedia

If any named entity with an entry in Wikipedia can be identified, then hypothesis on the likelihood of recognizing all Wikipedia articles on these entities can be reached. Therefore, the proposed classification algorithm is applied on the English Wikipedia dump dated third February 2014. Table 6 shows the number of each named entity type extracted from Wikipedia database. The number of named entities obtained through this approach (1575966) significantly outnumbers the figure of Wikipedia articles on named entities (1547586) derived from the same database in [8]. One may argue that this has been an earlier study while Wikipedia is constantly growing in size. This is true to an extent, however this study has only considered three types of named entities while [8] contains Miscellaneous named entities in addition to the three considered by this work. The generated database of named entities can be used as a training data for supervised classification strategies.

TABLE VI. SUMMARY OF EXTRACTED WIKIPEDIA NES

| Person | Location | Organization | total |
|---------------|----------|--------------|---------|
| 620790 | 290134 | 665042 | 1575966 |

V. CONCLUSION

A Wikipedia-based approach for predicting three types of named entities namely; Person, Location and Organization using article infoboxes is presented. Unlike common state of the art approaches which rather employ a set of multiple features such as article text, categories, links, among others, this study relies on a single feature consisting of the structured information in the infobox table. This has significantly reduced the classifier's processing time, which would be useful for delay sensitive applications requiring identification of designated names. Despite the use of a single feature, the proposed approach achieves a classification accuracy of above 97% with 3600 named entities and CoNLL-2003 shared task NER dataset used to validate the classifier's performance. Applying the same algorithm on Wikipedia database has resulted in the extraction of around 1.6 million named entities belonging to these three types. As a future work, the ongoing study aims to extend the infobox-based entity identification to generate a fine-grained entity classes in which each of the main types can be further subdivided into multiple subtypes.

REFERENCES

[1] R. Grishman and B. Sundheim, "Message Understanding Conference-6: A Brief History," in COLING, 1996, pp. 466-471.
[2] H. Liu and P. Singh, "ConceptNet—a practical commonsense reasoning tool-kit," BT technology journal, vol. 22, pp. 211-226, 2004.
[3] F. Wu and D. S. Weld, "Open information extraction using Wikipedia," in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 118-127.
[4] J. Liu and L. Birnbaum, "Measuring semantic similarity between named entities by searching the web directory," in Proceedings of the IEEE/WIC/ACM international Conference on Web intelligence, 2007, pp. 461-465.
[5] S. Cucerzan, "Large-Scale Named Entity Disambiguation Based on Wikipedia Data," in EMNLP-CoNLL, 2007, pp. 708-716.

[6] W. Zhang, J. Su, C. L. Tan, and W. T. Wang, "Entity linking leveraging: automatically generated annotation," in Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp. 1290-1298.
[7] J. Knopp, "Extending a multilingual Lexical Resource by bootstrapping Named Entity Classification using Wikipedia's Category System," Proceedings of the Fifth International Workshop On Cross Lingual Information Access, vol. 5, 2011, 35-43.
[8] W. Wentland, J. Knopp, C. Silberer, and M. Hartung, "Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration," in Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC, Marrakech, 2008.
[9] M. Tkachenko, A. Ulanov, and A. Simanovsky, "Fine Grained Classification of Named Entities In Wikipedia," HP Laboratories Technical Report-HPL-2010-166, 2010.
[10] P. Gamallo and M. Garcia, "A resource-based method for named entity extraction and classification," in Progress in Artificial Intelligence, ed: Springer, 2011, pp. 610-623.
[11] W. Dakka and S. Cucerzan, "Augmenting Wikipedia with Named Entity Tags," in IJCNLP, 2008, pp. 545-552.
[12] Y. Watanabe, M. Asahara, and Y. Matsumoto, "A Graph-Based Approach to Named Entity Categorization in Wikipedia Using Conditional Random Fields," in EMNLP-CoNLL, 2007, pp. 649-657.
[13] A. Bhole, B. Fortuna, M. Grobelnik, and D. Mladenić, "Extracting Named Entities and Relating Them over Time Based on Wikipedia," Informatica, vol. 31, 2007, 463-468.
[14] T. Zesch, I. Gurevych, and M. Mühlhäuser, "Analyzing and accessing Wikipedia as a lexical semantic resource," Data Structures for Linguistic Resources and Applications, pp. 197-205, 2007.
[15] Wikimedia. (2014). Wikipedia Statistics. Available: <https://stats.wikimedia.org/EN/Sitemap.htm#comparisons>
[16] P. Shachaf and N. Hara, "Beyond vandalism: Wikipedia trolls," Journal of Information Science, vol. 36, pp. 357-370, 2010.
[17] M. A. Hadj Taieb, M. Ben Aouicha, and A. Ben Hamadou, "Computing semantic relatedness using Wikipedia features," Knowledge-Based Systems, vol. 50, pp. 260-278, 2013.
[18] A. Bawakid, M. Oussalah, N. Afzal, S.-O. Shim, and S. Ahsan, "Disambiguating Words Senses with the Aid of Wikipedia," Life Science Journal, vol. 10, 2013, 1414-1426.
[19] P. Wang, J. Hu, H.-J. Zeng, and Z. Chen, "Using Wikipedia knowledge to improve text classification," Knowledge and Information Systems, vol. 19, pp. 265-281, 2009.
[20] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia," Artificial Intelligence, vol. 194, pp. 28-61, 2013.
[21] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from Wikipedia," Artificial Intelligence, vol. 194, pp. 151-175, 2013.
[22] J. Nothman, J. R. Curran, and T. Murphy, "Transforming Wikipedia into named entity training data," in Proceedings of the Australian Language Technology Workshop, 2008, pp. 124-132.
[23] D. Lange, C. Böhm, and F. Naumann, "Extracting structured information from Wikipedia articles to populate infoboxes," in Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 1661-1664.
[24] B. C. Ed Summers. (2011). WWW::Wikipedia - Automated interface to the Wikipedia. Available: <http://search.cpan.org/~bricas/WWW-Wikipedia-2.01/>
[25] T. Riddle, "Parse::MediaWikiDump- Tools to process MediaWiki dump files," 2010.
[26] A. Sultana, Q. M. Hasan, A. K. Biswas, S. Das, H. Rahman, C. Ding, and C. Li, "Infobox suggestion for Wikipedia entities," in Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, pp. 2307-2310.
[27] S. Tardif, J. R. Curran, and T. Murphy, "Improved text categorisation for Wikipedia named entities," in Australasian Language Technology Association Workshop 2009, 2009, p. 104-109.

Design and Implementation of an Interpreter Using Software Engineering Concepts

Fan Wu

Department of Computer Science
Tuskegee University
Tuskegee, Alabama, USA

Hira Narang

Department of Computer Science
Tuskegee University
Tuskegee, Alabama, USA

Miguel Cabral

Department of Computer Science
Tuskegee University
Tuskegee, Alabama, USA

Abstract—In this paper, an interpreter design and implementation for a small subset of C Language using software engineering concepts are presented. This paper reinforces an argument for the application of software engineering concepts in the area of interpreter design but it also focuses on the relevance of the paper to undergraduate computer science curricula. The design and development of the interpreter is also important to software engineering. Some of its components form the basis for different engineering tools. This paper also demonstrates that some of the standard software engineering concepts such as object-oriented design, design patterns, UML diagrams, etc., can provide a useful track of the evolution of an interpreter, as well as enhancing confidence in its correctness

Keywords—Interpreter; Software Engineering; Computer Science Curricula

I. INTRODUCTION

In this paper, an interpreter design and implementation for a small subset of C Language using software engineering concepts are presented. This paper summarizes the development process used, detail its application to the programming language to be implemented and present a number of metrics that describe the evolution of the project. Incremental development is used as the software engineering approach because it interleaves the activities of specification, development, and validation. The system was developed as a series of versions (increments) where each version adds functionality to the previous version [1].

The paper will also focus on the relevance of compilers and interpreters to undergraduate computer science curricula, particularly at Tuskegee University. Interpreters and compilers represent two traditional but fundamentally different approaches to implementing programming languages. A correct understanding of the basic mechanisms of each is an indispensable part of the knowledge that every computer science student must acquire [2].

The paper is organized as follows: section II presents the background and related work, and section III describes the design and development process. The conclusions and future work are discussed in section IV.

II. BACKGROUND AND RELATED WORK

A. Background

The main purpose of a compiler or an interpreter is to translate a source program written in a high-level source language to machine language. The language used to write the compiler or interpreter is called implementation language. The difference between a compiler and an interpreter is that a compiler generates object code written in the machine language and the interpreter executes the instructions. A utility program called a linker combines the contents of one or more object files along with any needed runtime library routines into a single object program that the computer can load and execute. An interpreter does not generate an object program. When you feed a source program into an interpreter, it takes over to check and execute the program. Since the interpreter is in control when it is executing the source program, when it encounters an error it can stop and display a message containing the line number of the offending statement and the name of the variable. It can even prompt the user for some corrective action before resuming execution of the program.

The process is divided into 6 functional increments. Before moving to the next increment, the current increment has to be tested and validated. The increments are: 1. the framework, 2. the scanner, 3. the symbol table, 4. parsing and interpreting expressions and assignment statements, 5. parsing and interpreting control statements, 6. parsing and interpreting declarations.

Interpreters are complex programs, and writing them successfully is hard work. To tackle the complexity, a strong software engineering approach can be used. Design patterns, Unified Modeling Languages (UML) diagrams, and other modern object-oriented design practices make the code understandable and manageable [3, 4, 5].

B. Related Work

While the area of interpreter design, as a subset of compiler design is well-established and documented, it is not typically the subject of formalized software engineering concepts.

The application of object-oriented design principles to parsers and compilers has been investigated by Reiss and Davis [6]. Malloy, Power, and Waldon reinforce the argument for the application of software engineering concepts in the area of parser design [7]. Similarly, an incremental approach to compiler design is proposed by Ghuloum [8].

Demille [9] states that compiler construction is a challenging process that requires material from virtually all computer science courses on the core curriculum. While the idea of compilers is usually furthered and explored in detail later on in an upper level course such as Compiler Construction, Xing [2] argues that the idea of interpreters rarely gets the same “treatment”: There is no such a course targeting on interpreter constructions in most undergraduate computer science curricula at universities and colleges [9].

III. DESIGN AND IMPLMENTATION

A. Conceptual Design

The conceptual design of a program is a high-level view of its software architecture. The conceptual design includes the primary components of the program, how they're organized, and how they interact with each other. An interpreter is classified as a programming language translator. A translator, as seen at the highest level, consists of a front end and a back end. Both compilers and interpreters can share the same front end, but they'll have a different back end. Fig. 1 shows the conceptual design of the SimpleC interpreter. The front end of a translator reads the source program and performs the initial translation stage. Its primary components are the parser, the scanner, the token, and the source.

The parser controls the translation process in the front end. It continuously asks the scanner for the next token, and it analyzes the sequences of tokens to determine what high-level language elements it is translating. The parser verifies that what it sees is syntactically correct as written in the source program; in other words, the parser detects and flags any syntax errors. The scanner reads the characters of the source program sequentially and constructs tokens, which are the low-level elements of the source language. The scanner scans the source program to break it apart into tokens.

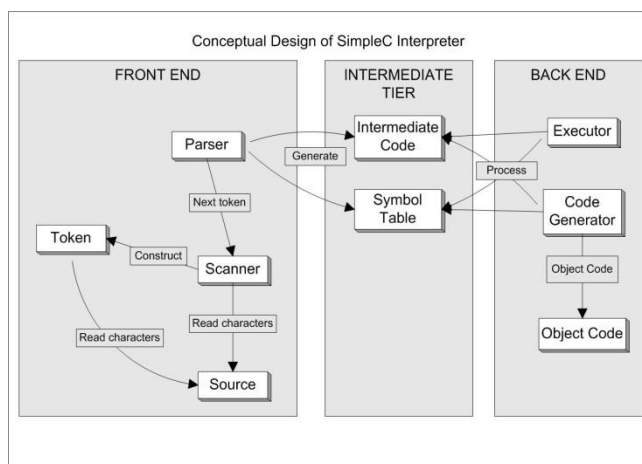


Fig. 1. Conceptual Design of the SimpleC Interpreter/Compiler

1) Syntax and Semantics

The syntax of a programming language is its set of grammar rules that determine whether a statement or an expression is correctly written in that language. The language's semantics give meaning to a statement or an expression. In Simple C, the statement (1):

$$a = b + c; \quad (1)$$

is a valid assignment statement. The semantics of the language tells that the statement says to add the value of variables 'b' and 'c' and assign the sum's value to the variable 'a'.

A parser performs actions based on both the source language's syntax and semantics. Scanning the source program and extracting tokens are syntactic actions. Looking for '=' token is a syntactic action, entering the identifiers 'a', 'b', and 'c' into the symbol table as variables, or looking them up in the symbol table, are semantic actions because the parser had to understand the meaning of the expression and the assignment to know that it needs to use the symbol table. Syntactic actions occur in the front end, while semantic actions can occur on either the front end or the back end.

B. Basic Interpreter/Compiler Framework

As mentioned previously, the project will be divided into functional increments, using software engineering concepts. In the previous section, the conceptual design of the compiler was briefly explained. In this increment, an initial implementation of a rudimentary interpreter will be presented after conceptual design.

The first part of this increment is to build a flexible framework that supports both compilers and interpreters. The framework will integrate fundamental interpreter and compiler components in the second stage. Finally, end-to-end tests will be run to test the framework and its components.

The goals for this increment are:

- A source language-independent framework that can support both compilers and interpreters
- Initial SimpleC source language-specific components integrated into the front end of the framework
- Initial compiler and interpreter components integrated into the back end of the framework
- Simple end-to-end runs that exercise the components by generating source program listings from the common front end and messages from the compiler or interpreter back end

1) Front End

The front end consists of the language-independent *Parser*, *Scanner*, *Source*, and *Token* classes that represent the framework's components. Consider the class diagram of the front end in Fig. 2.

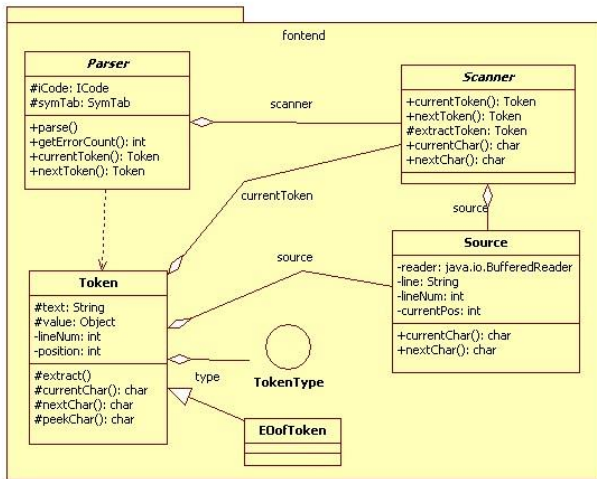


Fig. 2. Front end package

The parser and the scanner are closely related. The parser “owns” a scanner. The parser request tokens from its scanner, and so it has a dependency on tokens. The scanner owns the current token, it owns the source, and it passes the source reference to each token it constructs. Each token then also owns that. During its construction, a token reads characters from the source.

2) Messages

The parser may need to report some status information, such as an error message whenever it finds a syntax error. However, the parser should not worry about where it should send the message or what the recipient does with it. Similarly, whenever the source component reads a new line, it can send a message containing the text of the line and the line number. Keeping the senders of messages loosely coupled to the recipients of the messages minimize their dependencies. In complex applications, loose coupling allows you to develop components independently and in parallel [7].

3) Intermediate Tier

According to the conceptual design, the intermediate code and the symbol table are the interface between the front and back ends. Consider the following UML class diagram.

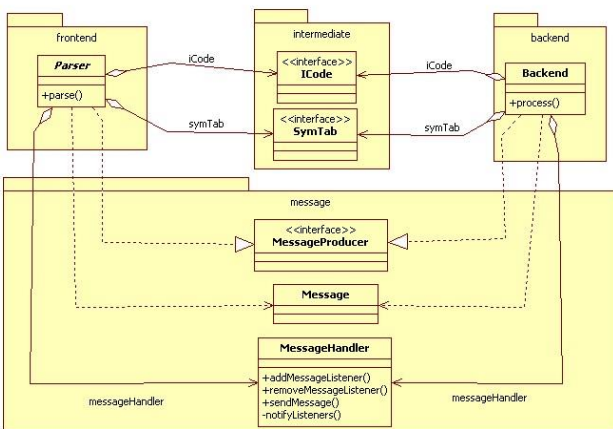


Fig. 3. Front end, intermediate, and backend packages

A framework Parser object in the front end package and a Backend object in the backend package own intermediate code and symbol table object as shown in Fig.3. Both classes, Parser and Backend, have the same relationships to the classes in the message package.

4) Back End

The conceptual design states that the back end will support either a compiler or an interpreter. Just like the Parser and Source classes in package front end, the Back End class in package backend implements the MessageHandler helper class. A compiler would implement the abstract method process to generate object code while an interpreter would implement the same method to execute the program.

5) Test and Validation

Up to this point a framework with components that are language independent has been completed. The backend of the framework can support either a compiler or an interpreter. To test the framework, a SimpleC program is used as the input. The output consists on the input program with line numbers followed by the number of statements, number of syntax errors, total parsing time, statements executed, run-time errors, and total execution time. The error recovery will be added in next increments.

C. Scanning

The second increment of the project consists on implementing a scanner. The scanner is the component in the front end of a compiler or an interpreter that performs the syntactic actions of reading the source program and breaking it apart into tokens. The parser calls the scanner each time it wants the next token from the source program. The goals for this increment are:

- Complete the design and development of the SimpleC scanner.
- The scanner should be able to:
- Extract SimpleC words, numbers, and special symbols from the source program
- Determine whether a word is an identifier or a SimpleC reserved word
- Calculate the value of a number token and determine whether its type is integer or real
- Perform syntax error handling

1) Syntax Error Handling

Every parser must be able to handle syntax errors in the source program. Error handling is a three step process:

- Detection: Detect the presence of a syntax error.
- Flagging: Flag the error by pointing it out or highlighting it, and display a descriptive error message.
- Recovery: Move past the error and resume parsing.

If an instance of one of the SimpleC token subclasses finds a syntax error, it will set its type field to the SimpleCTokenType enumerated value ERROR and its value field to the appropriate SimpleCErrorCode enumerated value.

If the scanner finds syntax errors (such as an invalid character that cannot start a legitimate SimpleC token), it will construct a SimpleCErrorToken.

2) How to Scan for Tokens

The scanner has to read each character at a time, skipping blank spaces. For example consider the following statement:

```
int a = 3; (2)
```

After scanning the statement, the scanner has extracted the following tokens:

| TYPE | TEXT STRING |
|----------------------|-------------|
| Word (reserved word) | int |
| Word (identifier) | a |
| Special symbol | = |
| Number (integer) | 3 |
| Special symbol | ; |

The scanner reads and skips white space characters between tokens. When it's done, the current character is nonblank. This nonblank character determines the type of the token the scanner will extract next, and the character becomes the first character of that token. The scanner extracts a token by reading and copying successive source characters up to but not including the first character that cannot be part of the token. Extracting a token consumes all the source characters that constitute the token. Therefore, after extracting a token, the current character is the first character after the last token character. The SimpleC scanner can identify word tokens (identifiers and reserved words), special symbol tokens ('+', '-', etc.), and number tokens (unsigned integers and real numbers).

3) Test and Validation

To test this increment a SimpleC tokenizer utility was written. The tokenizer takes as input a SimpleC source program and outputs a description of a token or an error message in case of a syntax error. Fig. 4 shows the output of the tokenizer for input file simplec_mult.txt.

```
001 // file-name simplec_mult.txt
002 .
>>> DOT line=002, pos = 0, text="."
003 int mult = 2 * 4;
>>> INT line=003, pos = 0, text="int"
>>> IDENTIFIER line=003, pos = 2, text="mult"
>>> ASSIGN line=002, pos = 3, text="="
>>> INTEGER line=002, pos = 4, text="2"
>>> STAR line=002, pos = 5, text="*"
>>> INTEGER line=002, pos = 6, text="4"
004 .
>>> DOT line=004, pos = 0, text="."
```

Fig. 4. Output of SimpleC Tokenizer

D. The Symbol Table

The parser of a compiler or an interpreter builds and maintains a symbol table throughout the translation process as part of semantic analysis. The symbol table stores information about the source program's tokens, mostly the identifiers. As mentioned in previous increments, the symbol table is a key component in the interface between the front and back end.

Goals for this increment:

- A language-independent symbol table
- A simple utility program that parses a SimpleC source program and generates a cross-reference listing of its identifiers

1) The Symbol Table

The approach of this increment is to create the conceptual design of a symbol table, develop interfaces that represent the design, and finally write the classes that implement the interfaces. To verify the correctness of the source code, a cross-reference utility program will be used. It will exercise the symbol table by entering, finding, and updating information.

During the translation process, the interpreter creates and updates entries in the symbol table to contain information about certain tokens in the source program. Each entry has a name, which is the token's text string. The entry also contains information about the identifier. As it translates the source program, the interpreter looks up and updates the information.

The symbol table entry for an identifier will typically include its type, structure, and how it was defined. One of the goals is to keep the symbol table flexible and not limited to SimpleC-specific information. The basic operations a symbol table must support are: 1. Enter new information, 2. Loop up existing information, 3. Update existing information.

2) Conceptual Design of Symbol Table Stack

To parse a language like SimpleC, more than one symbol table might be needed (one table for each function or class, etc.). Because some procedures can be nested, the symbols need to be maintained in a stack. Fig 5 shows the conceptual design of the symbol stack. For this increment, only one table will be used but the concept will be explained for possible future extensions of the language. The symbol table at the top of the stack maintains information for the program, function, block, etc. that the parser is currently working on. As the parser works its way through the SimpleC program and enters and leaves nested functions and blocks, it pushes and pops symbol tables from the stack. The symbol table at the top of the stack is known as the local table.

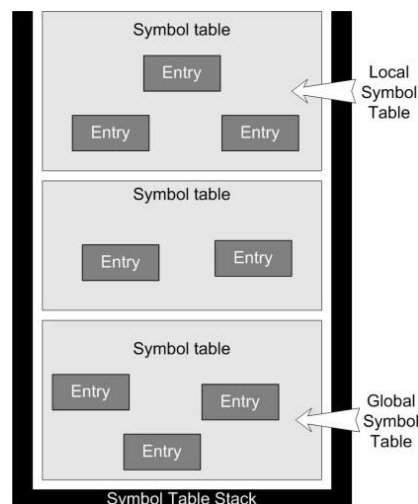


Fig. 5. The conceptual design of the symbol stack.

3) Test and Validation

The last step of the increment is to test the symbol table. This goal can be accomplished by generating a cross-reference of a SimpleC source program. The command line:

```
execute -x example01.txt
```

is used to generate a cross-reference listing of the identifiers found in the file "example01.txt".

```
----jGRASP exec: java SimpleC compile -x simplec_assign_ex.txt

001 // file-name: simplec_assign_ex.txt
002 .
003 {
004 five = 5;
005 ten = 10;
006 fifteen = five + ten;
007 }
008 .
009

          9 source lines.
          0 syntax errors.
          0.03 seconds total parsing time.

===== CROSS-REFERENCE TABLE =====

Identifier      Line numbers
-----
fifteen         006
five            004 006
ten             005 006

          0 instructions generated.
          0.00 seconds total code generation time.

----jGRASP: operation complete.
```

Fig. 6. Cross-reference table for simplec_assign_ex.txt

After the source program listing, all of the source program's identifiers are listed alphabetically. Following each identifier name are the source line numbers where the identifier appears as shown on Fig. 6.

E. Expressions and Assignment Statements

In the previous increment, a symbol table was created. The parser builds and maintains the symbol tables on the symbol table stack during the translation process. The parser also performs the semantic actions of building and maintaining intermediate code that represents the source program in the form of parse trees. The back end will then interpret the parse trees in order to execute statements and expressions. The goals for this increment are:

- Parsers in the front end for certain SimpleC constructs: assignment statements, compound statements, and expressions.
- Flexible, language-independent intermediate code generated by the parsers to represent these constructs.
- Language-independent executors in the interpreter back end that will interpret the intermediate code and execute expressions and assignment statements.

1) Syntax Diagrams

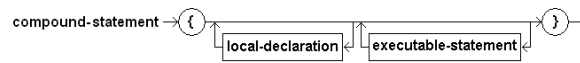
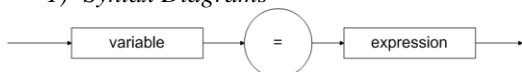


Fig. 7. Assignment statement and compound statement

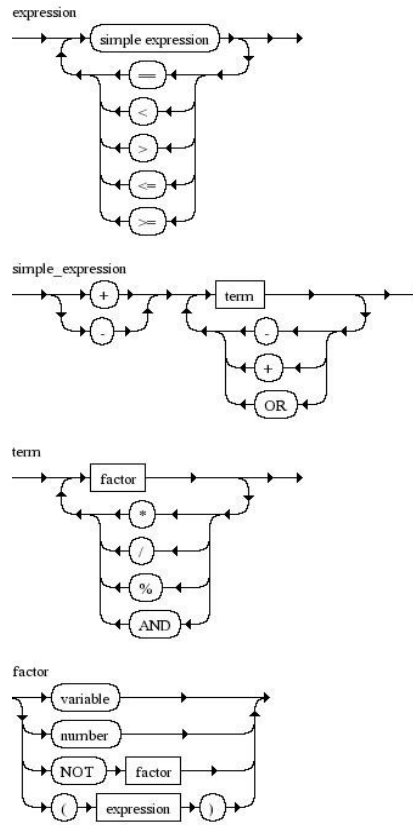


Fig. 8. Syntax diagrams for SimpleC expressions

Figs. 7 and 8 show the syntax diagrams that guide the development of the parsers that will generate the appropriate intermediate code.

2) Intermediate Code

A data tree structure represents the SimpleC intermediate code. Therefore, the intermediate code takes the form of a parse tree. A parse tree consists of sub-trees that represent SimpleC constructs, such as statements and expressions. Each tree node has a node type and a set of attributes. Each node other than the root node has a single parent node. The industry-standard XML can represent the tree structures in text form.

3) Executing Expressions and Assignment Statements

Expressions and statements are executed in the back end of the interpreter.

The intermediate code that represents the parse trees was implemented using language-independent classes in the back end. Consider the UML diagram if Fig. 9 for the statement executor classes in the back end package.

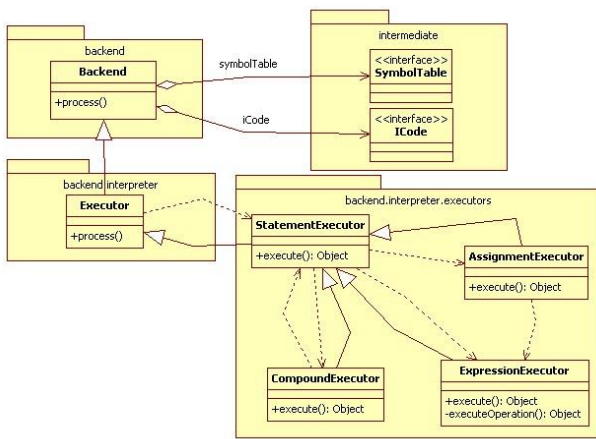


Fig. 9. Executor subclasses in the back end

4) Test and Validation

To test and validate this increment, a simple interpreter was written. The simple interpreter takes as input a SimpleC source program. Only assignment statements, compound statements, and expressions are recognized by the interpreter. Fig. 10 shows the output of file simplec_assign.txt.

```

----jGRASP exec: java SimpleC execute simplec_assign_ex.txt
001 // file-name: simplec_assign_ex.txt
002 .
003 {
004 a = 10;
005 b = a + 2;
006 c = 2 * (b - 2);
007 }
008 .
009

          9 source lines.
          0 syntax errors.
          0.03 seconds total parsing time.

----- OUTPUT -----
>>> LINE 004: a = 10
>>> LINE 005: b = 12
>>> LINE 006: c = 20

          3 statements executed.
          0 runtime errors.
          0.00 seconds total execution time.

----jGRASP: operation complete.
    
```

Fig. 10. Output of simplec_assign.txt

F. Control Statements

The next increment focuses on parsing and interpreting control statements. The goals for this increment are:

- Parsers in the front end for SimpleC control statements if, while, and for.
- Flexible, language-independent intermediate code generated by the parsers to represent these constructs.
- Reliable error recovery to ensure that the parsers can continue to work despite syntax errors in the source program.

The syntax diagrams shown in Fig. 11 were used to guide the development of the parsers.

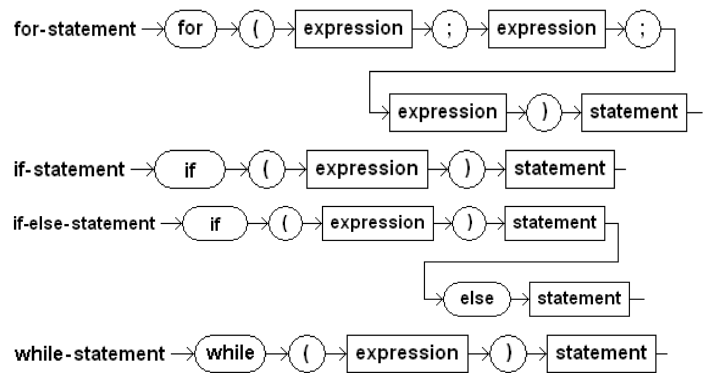


Fig. 11. Syntax diagrams for SimpleC control statements

1) Error Recovery

When the parser encounters an error, the three possible options for error recovery are: 1. Terminate the program after encountering a syntax error, 2. Attempt to parse the rest of the source program, 3. Skip tokens after the erroneous one until it finds a token it recognizes and safely resume syntax checking.

The first two options are undesirable. To implement the third option, a parser must “synchronize” itself frequently at tokens it expects. Whenever there is a syntax error, the parser must find the next token in the source program where it can reliably resume syntax checking [7].

2) Interpreting Control Statements

The interpreting capabilities of the program increase after each increment. It is time to add new executor classes for SimpleC control statements. The control statement executor classes can be appreciated in Fig. 12.

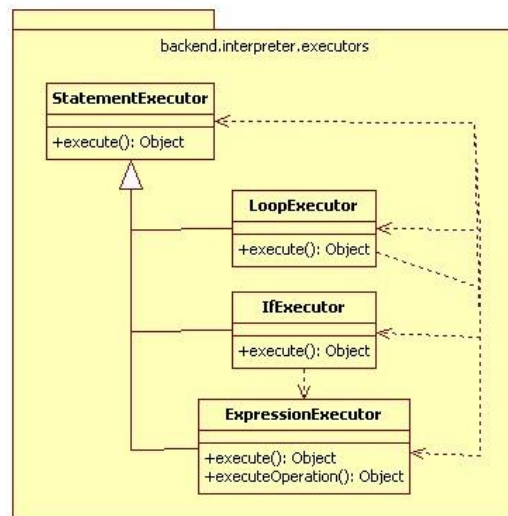


Fig. 12. Control statement executors in the backend

3) Test and Validation

To test this increment, a syntax checker utility was written to identify syntax errors. Also, the simple interpreter described in the previous increment was expanded. The interpreter takes as input a SimpleC source program. At this point, the program

identifies conditional statements and loop statements. Fig. 13 shows a sample output of a SimpleC if-statement.

```

001 // file-name simplec_if.txt
002 .
003 a = 2;
004 b = 3;
005 if(b > a)
006 {
007   a = a + b;
008 }
009 .

      8 source lines.
      0 syntax errors.
      0.02 seconds total parsing time.

----- OUTPUT -----
>>> LINE 003: a = 2
>>> LINE 004: b = 3
>>> LINE 007: a = 5
      4 statements executed.
      0 runtime errors.
      0.01 total execution time.
    
```

Fig. 13. Execution of a SimpleC if statement

G. Parsing Declarations

Parsing declarations expands the work in The Symbol Table increment because all the information from the declarations has to be entered in the symbol table. The goals for this increment are:

- Parsers in the front end for SimpleC type definitions and type specifications.
- Additions to the symbol table to contain type information.

1) SimpleC Declarations

There are three basic types of variables in SimpleC; they are: char, int, and float. The syntax diagram is shown in fig. 14.

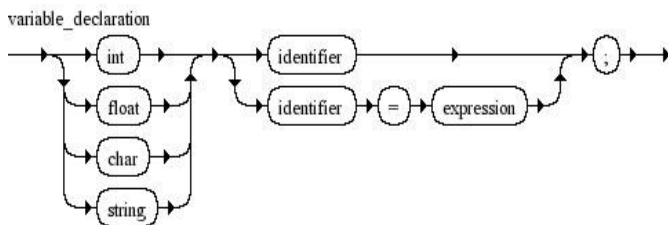


Fig. 14. Syntax diagram for variable declarations

2) Types and the Symbol Table

The type specification parser developed in this increment enters type information into the symbol table. The first step is to design language-independent interfaces that treat a type specification simply as a collection of attributes.

3) Parsing SimpleC Declarations

In previous increments it is assumed that identifiers were variables. For this increment, an identifier's symbol table entry must indicate how it was defined. Fig. 15 shows the classes for parsing declarations.

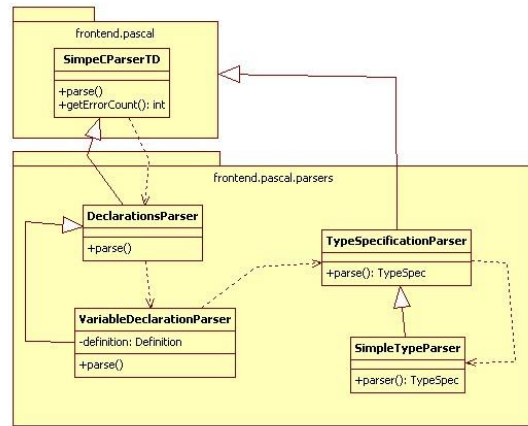


Fig. 15. The classes for parsing declarations

4) Test and Validation

To test and validate the code written for this increment, a SimpleC cross-reference utility similar to the one implemented in The Symbol Table increment, is implemented. The output of the cross-reference utility includes the line number where an identifier is found and how it is defined, as shown in Fig. 16.

| ==== CROSS-REFERENCE TABLE ==== | | |
|------------------------------------|--------------|---------------------|
| *** PROGRAM simplec_assign.txt *** | | |
| Identifier | Line numbers | Type specification |
| a | 002 005 | Defined as: integer |
| b | 003 006 | Defined as: integer |
| c | 004 007 | Defined as: real |
| ... | | |

Fig. 16. Sample output of cross-reference table

IV. CONCLUSIONS AND FUTURE WORK

A. Conclusions

In this paper, the design of an interpreter for the SimpleC programming language in the context of a software engineering project has been presented. The paper also has demonstrated that some of the standard software engineering concepts such as object-oriented design, design patterns, UML diagrams, etc., can provide a useful track of the evolution of an interpreter, as well as enhancing confidence in its correctness. A similar project could be introduced at Tuskegee University to meet some requirements not satisfied by shorter projects. Some requirements include, but are not limited to, writing a complete project using challenging algorithms and data structures, use of different development tools, object-oriented design, and team management which is an important issue to consider given that only team work in software engineering and database courses.

B. Future Work

Future work will focus on creating an interactive source-level debugger for the SimpleC language that enables the use of command lines to interact with the interpreter as well as an Integrated Development Environment (IDE) with a graphical user interface (GUI). If time is not a constraint, the interpreter will be extended to a SimpleC compiler that generates object

code for the Java Virtual Machine (JVM). The compiled programs will then be able to run on multiple platforms.

REFERENCES

- [1] Sommerville, I. *Software Engineering*. Addison Wesley, 9th edition, 2010
- [2] C. Xing. "How Interpreters Work: An Overlooked Topic in Undergraduate Computer Science Education," Proc. In CCSC Southern Eastern Conference, JCSC Vol. 25, Issue 2. December 2009
- [3] R. Mak. *Writing Compilers and Interpreters: A Modern Software Engineering Approach Using Java*. Wiley, 3rd edition, 2009
- [4] H. Deitel, *Java: How to Program (early projects)*, 10th edition, Prentice Hall Inc. , 2014.
- [5] R. Sebesta, *Concepts of Programming Languages*, 10th Edition, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 2012.
- [6] S. Reiss, and T. Davis. "Experiences Writing Object-Oriented Compiler Front Ends". Tech. Rep., Brown University, January 1995.
- [7] B. Malloy., J. Power, and J. Waldron. "Applying Software Engineering Techniques to Parser Design: The Development of a C# Parser," in Proc. of SAICSIT 2002, pp. 74–81
- [8] A. Ghuloum, "An Incremental Approach to Compiler Construction," in Proc. of the 2006 Scheme and Functional Programming Workshop, 2006, pp. 28.
- [9] A. Demaille. "Making Compiler Construction Projects Relevant to Core Curriculums," In proceeding of: Proceedings of the 10th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, ITiCSE 2005, Caparica, Portugal, June 27-29, 2005.

A parallel line sieve for the GNFS Algorithm

Sameh Daoud

Computer Science Division,
Department of Mathematics

Faculty of Science, Ain Shams University, Cairo, Egypt
Email: Sameh_Daoud2003@yahoo.com

Ibrahim Gad

Computer Science Division,
Department of Mathematics

Faculty of Science, Tanta University, Tanta, Egypt
Email: gad_12006@yahoo.com

Abstract—RSA is one of the most important public key cryptosystems for information security. The security of RSA depends on Integer factorization problem, it relies on the difficulty of factoring large integers. Much research has gone into problem of factoring a large number. Due to advances in factoring algorithms and advances in computing hardware the size of the number that can be factorized increases exponentially year by year. The General Number Field Sieve algorithm (GNFS) is currently the best known method for factoring large numbers over than 110 digits. In this paper, a parallel GNFS implementation on a BA-cluster is presented. This study begins with a discussion of the serial algorithm in general and covers the five steps of the algorithm. Moreover, this approach discusses the parallel algorithm for the sieving step. The experimental results have shown that the algorithm has achieved a good speedup and can be used for factoring a large integers.

Keywords—parallel Algorithm; Public Key Cryptosystem; GNFS Algorithm.

I. INTRODUCTION

Factoring is very important in the field of cryptography, specifically in the RSA cryptosystem. The RSA algorithm [5] is the most popular algorithm in public-key cryptosystems and RSA is used in real world applications such as: internet explorer, email systems, and online banking [12]. The security of RSA algorithm relies on the difficulty of factoring large integers. There are many integer factorization algorithms used to factor large numbers, such as Trial division [6], Pollards p-1 algorithm [7], Lenstra Elliptic Curve Factorization (ECM) [8], Quadratic Sieve (QS) [9] and General Number Field Sieve (GNFS) algorithm [1]–[4]. GNFS is the best known algorithm for factoring large composite numbers over than 110 digits. This algorithm takes a long time to factor large integers. Therefore, this paper presents an implementation of parallel GNFS algorithm on a BA-cluster.

The main objective of this paper giving new proposed algorithm for sieving step in cluster system. This paper consists of eight sections. Section II will introduce the GNFS algorithm. Section III gives the reasons for selecting sieve step. Section IV gives an overview for serial sieve step and give an overview for previous parallel sieve step. Section V proposes a new method for parallel sieve step on cluster system. In section VI the configuration of hardware and software used to implement the parallel sieving step on cluster system. Section VII introduces the experimental results for the proposed methods. Section VIII focus in conclusion and future works.

II. THE GNFS ALGORITHM

The General Number Field Sieve (GNFS) algorithm [1], [2] is derived from the Number Fields Sieve (NFS) algorithm, developed by A. K. Lenstra, H. W. Lenstra, M.S. Manasse and J. M. Pollard [10].

GNFS have five major steps which are described as follows:

1) Step 1: (Polynomial selection)

Find a polynomial $f : \mathbb{R} \rightarrow \mathbb{R}$ of degree d with integer coefficients as follows:

$$f(x) = a_d x^d + a_{d-1} x^{d-1} + \dots + a_0 \quad (1)$$

such that $f(m) \equiv 0 \pmod{n}$.

2) Step 2: (Factor bases)

The main objective of this step is to find three types of factor bases: (1) rational factor base, R , (2) algebraic factor base, A , and (3) quadratic character base, Q . The three factor bases are define as follows:

The rational factor base R [1]

A rational factor base is a finite collection of prime numbers, p_i , up to some bound M , $M \in \mathbb{N}$. i.e.

$$R = \{p : p \text{ is a prime and } p \leq M, M \in \mathbb{N}\}.$$

Smooth over R [1]

An integer $l \in \mathbb{Z}$ is said to be smooth over a rational factor base R if R contains all of the prime divisors of l . i.e.

$$l = \prod_{p_i \in R} p_i.$$

The algebraic factor base A [1]

An algebraic factor base is a finite set $\{a+b\theta\} \subset \mathbb{Z}[\theta]$ where for $a, b \in \mathbb{Z}$, each $a+b\theta$ satisfies $\forall(a, b), \nexists c, d \in \mathbb{Z}[\theta]$ such that $c \cdot d = a + b\theta$.

Smooth over A [1]

An element $l \in \mathbb{Z}[\theta]$ is said to be smooth over an algebraic factor base A if

$$l = \prod_{(c,d) \in A \subset A} (c + d\theta).$$

The quadratic character base Q [1]

The quadratic character base Q is a finite set of pairs (p, r) with the same properties as the elements of the algebraic factor base, but the primes $p_i \in Q$ are larger than the largest in the algebraic factor base, $p_i > \hat{p} \in A$ where \hat{p} is the largest element in the algebraic factor base A , i.e.

| Digits | Sieve | Total | Sieve/Total |
|--------|----------|----------|-------------|
| 30 | 26.5s | 32.3s | 82% |
| 39 | 15.0s | 19.7s | 76.1% |
| 45 | 184.3s | 250s | 74% |
| 51 | 222.3s | 311.5s | 71.4% |
| 61 | 3620.7s | 4320.4s | 84% |
| 76 | 26477.9s | 37171.9s | 71.2% |
| 98 | 17300.6s | 20790.9s | 83.2% |

TABLE I: GNFS integer factorization records [14]

$$Q = \{(p_i, r_i) : p_i > \hat{p} \text{ where } \hat{p} \text{ the largest in the algebraic factor base } A\}$$

3) **Step 3: (Sieving)**

Find many pairs of integers (a, b) with the following properties:

1. $\gcd(a, b) = 1$.
2. $a + bm$ is smooth over the rational factor base.
3. $a + b\theta$ is smooth over the algebraic factor base.

4) **Step 4: (Linear algebra)**

The relations are put into relation sets and a very large sparse matrix is constructed. The matrix is reduced resulting in some dependencies, i.e. elements which lead to a square modulo n .

5) **Step 5: (Square root)**

- Calculate the rational square root, r , such that:

$$r^2 = \prod_{(a,b) \in V} (a + bm)$$

- Calculate the algebraic square root, s , such that:

$$s^2 = \prod_{(a,b) \in V} (a + b\theta)$$

- Then p and q can then be found by $\gcd(n, s - r)$ and $\gcd(n, s + r)$ where p and q are the factors of n .

III. WHY SIEVING STEP?

The main objective of this section is to give the importance of the sieving step. Previous studies shows that the sieving step is very important for several reasons:

- 1) The sieving step is the most time consuming, it takes more than 70% of the total time from the time of implementation as shown in Table I [14].
- 2) The second reason is that the sieving step can be parallelized easily.

The experimental studies show that there are some problems in the implementation that led to slow the previous parallel program. In the previous algorithm there are many communications between the master nodes and the slaves. The communication times increase when the size of n increases. Another cause for inefficiency is that each processor does sieving for different pairs. Therefore, the sieving time for each processor might be different. The master node can not start the next sieving until all the slave nodes finish their sieving [14].

IV. PREVIOUS SIEVING WORK

Algorithm 1 shows the steps of serial sieving. The sieving step uses nested for-loops, one for the values of b 's and the other for the values of a 's. In the outer loop, b ranges from $-C$ to C , usually the values of b 's are in range $1 \leq b < C$. In the inner loop, b is fixed and a changes from $-N$ to N . The sieving step takes long time because it uses two loops and the values of a and b are usually very large.

Algorithm 1 Serial sieving algorithm [14].

```

1:  $b_0 = 1$ ;
2:  $b_1 = C$ ;
3:  $a_1 = -N$ ;
4:  $a_2 = N$ ;
5: for ( $b = b_0; b < b_1; b++$ ) do
6:   for ( $a = a_1; a < a_2; a++$ ) do
7:     if ( $(a, b)$  Smooth over  $R$  and Smooth over  $A$ ) then
8:        $save(a, b)$ ;
9:     end if
10:   end for
11: end for

```

L.T.Yang, L.Xu, and M.Lin proposed parallel sieving in a cluster system [12]–[15]. The basic idea of the proposed algorithm is that each processor takes a range of b 's values and generate a set of (a, b) pairs as shown in algorithm 2.

Algorithm 2 Parallel sieving algorithm [14].

```

1: MPI_Init();
2: MPI_Comm_size();
3: MPI_Comm_rank();
4:  $b_0 = Min\_b$ ;
5:  $b_1 = Max\_b$ ;
6:  $a_1 = -N$ ;
7:  $a_2 = N$ ;
8:  $num\_of\_bs = ((b_1 - b_0) / p)$ ;
9: MPI_Bcast( $num\_of\_bs$ );
10: for ( $b = (taskid * num\_of\_bs + b_0) + 1; b \leq (b_0 + (taskid + 1) * num\_of\_bs; b++)$ ) do
11:   for ( $a = a_1; a \leq a_2; a++$ ) do
12:     if ( $(a, b)$  Smooth over  $R$  and Smooth over  $A$ ) then
13:       if master then
14:         MPI_Recv(( $a, b$ ));
15:          $save(a, b)$ ;
16:       else
17:         MPI_Send(( $a, b$ ));
18:       end if
19:     end if
20:   end for
21: end for
22: MPI_Finalize();

```

V. THE NEW METHODS

The main objective of this section is to describe the new methods for the parallel sieving step of GNFS algorithm. The new methods improve the parallel sieving algorithm by decreasing the communications between the master node and the slaves. In the following sections (V-A, V-B) we explain the new methods and the results of each method.

A. The first method

The main idea of the first method is to divide the range of b between the processors. This is because, each b in the outer loop generates a set of ordered pairs (a, b) independently of the others b 's. So, each processor takes a range of b values and generates a set of (a, b) pairs and then saves in a local file, see Fig. 1. When all processors finish the computations of finding their sets of ordered pairs (a, b) , the master node copy all the files that have the sets of (a, b) pairs from the slaves into one file.

B. The Second Method

The main idea of the second method is the same as the first method, it depends on dividing the range of b between the processors. Except that each processor takes a range of b values and generate a set of (a, b) pairs and then save it in an array of large size ($rels$), see Fig. 2. Then each slave, find the $rels$ of different sets of ordered pairs (a, b) for each b in the range belonging to this slave, then the slave will send the $rels$ to the master, and the master node receives all the sets of $rels$ from the slaves. This process will be repeated until we reach the last b in the range belonging to this slave.

VI. HARDWARE AND SOFTWARE PROGRAMMING ENVIRONMENT

The parallel GNFS program is implemented on a Bibliotheca Alexandrina (BA) Supercomputer which is located in Alexandria library, Alexandria, Egypt. The supercomputer is a high performance computing cluster with performance reaching 11.8 TFLOPS. It is composed of 130 computational nodes, 6 management nodes including two batch nodes for job submission (64 Gbyte RAM), inter-process Communication network, and 36-TByte storage. Each node has two Intel Quad core Xeon 2.83 GHz processors (64 bit technology), 8 Gbyte RAM, 80 Gbyte hard disk, and a GigaEthernet network port.

The parallel code is based on the serial code developed by C. Monico in [3]. The program is written in ANSI C and compiled by GNU C compiler (gcc) and run under Linux operating system. We have used MPI library to write the parallel program. MPICH1 [16] is installed for MPI library. Also we installed a free library GMP [17] which is required to compile and to run the program.

VII. PERFORMANCE EVALUATION

A. Test Cases

In order to test our parallel algorithm for speedup and efficiency, we choose different n and different number of processors. In Table II shows all test cases and number of processors which are used.

B. Timing Results

The time for the first method and the second method for each test case is shown in Fig.3.

The Fig.3 show that the ruining time decreases by increasing the number of processors. From Fig.3 the first

| Test | Digits of n | Number of processors |
|------|---------------|-----------------------------|
| 1 | 61 | 1, 2, 4, 8, 10, 12, ,14, 16 |
| 2 | 76 | 1, 2, 4, 8, 10, 12, ,14, 16 |
| 3 | 80 | 1, 2, 4, 8, 10, 12, ,14, 16 |
| 4 | 100 | 1, 2, 4, 8, 10, 12, ,14, 16 |
| 5 | 110 | 1, 2, 4, 8, 10, 12, ,14, 16 |
| 6 | 120 | 1, 2, 4, 8, 10, 12, ,14, 16 |
| 7 | 130 | 1, 2, 4, 8, 10, 12, ,14, 16 |

TABLE II: Test cases and number of processors

method is faster than the second method using small number of processors, otherwise the two methods are approximately equal.

C. Speed-Up

Speedup is defined by the following formula: $S_p = \frac{T_1}{T_p}$, where p is the number of processors, T_1 is the execution time of the sequential algorithm, and T_p is the execution time of the parallel algorithm with p processors. The speedup for the test cases using different number of processors for the first method and for the second method are presented in Fig.4.

From Fig.4 the second method is better than the first method when n is small number, the first method is better than the second method when n is large number.

D. Sieving Efficiency

Efficiency is defined as $E_p = \frac{S_p}{p} = \frac{T_1}{pT_p}$. It is a value, between zero and one. The sieving efficiency for each test case is shown in Fig.5.

From Fig.5 the first method is better than the second method using small number of processors, otherwise is approximately equal.

VIII. DISCUSSIONS

We propose two algorithms for sieving step in cluster system. The difference between them is that one generate a set of (a, b) pairs and then save it in local file for each processor, the other strategy is to generate a set of (a, b) pairs and then save it in an array of large size then send the set of (a, b) pairs to the master.

The experimental studies show that the ruining time decreases by increasing the number of processors. From Fig.3 the first method is faster than the second method when using small number of processors, otherwise they are approximately equal. Fig.4 shows that the speed-up for the second method is better than the first method when n is small number, the first method is better than the second method when n is large number. Fig.5 shows that the efficiency for the first method is better than the second method using small number of processors, otherwise the efficiency is approximately equal.

There are still open questions and some research points which can be studied, in future, such as:

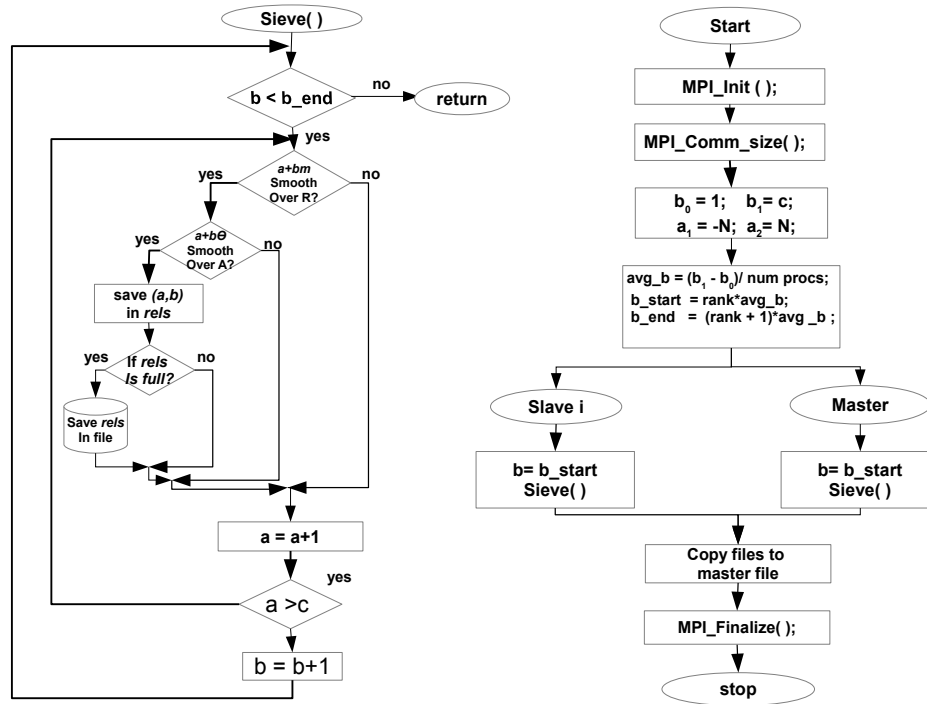


Fig. 1: The flowchart of first method

- 1) Decreasing the communications time when the size of n increases by decreasing the communications between the master nodes and the slaves, so the sieving time decreases when the communications decreases.
- 2) Further improvements on better load balance.
- 3) Trying to make all the steps of the algorithm in parallel whenever possible

REFERENCES

[1] M.Case, *A beginners guide to the general number field sieve*, pages 14, 2003

[2] M.E.Briggs. *An introduction to the general number field sieve*. Masters thesis, Virginia Polytechnic Institute and State University, 1998.

[3] C.Monico, *General number field sieve documentation.*, GGNFS Documentation, Nov 2004

[4] J.Dreibellbis., *Implementing the general number field sieve.* pages 5–14, June 2003

[5] R.L.Rivest, A.Shamir, and L.M.Adelman, *A method for obtaining digital signatures and public-key cryptosystems.* Technical Report MIT/LCS/TM-82, 1977.

[6] M.C.Wunderlich and J.L.Selfridge. *A design for a number theory package with an optimized trial division routine.* Communications of ACM, 17(5):272–276, 1974.

[7] J.M.Pollard. *Theorems on factorization and primality testing.* In Proceedings of the Cambridge Philosophical Society, pages 521–528, 1974.

[8] H.W.Lenstra. *Factoring integers with elliptic curves.* Annals of Mathematics(2), 126:649–673, 1987.

[9] C.Pomerance. *The quadratic sieve factoring algorithm.* In Proceeding of the EUROCRYPT 84 Workshop on Advances in Cryptology: Theory and Applications of Cryptographic Techniques, pages 169–182. Springer-Verlag, 1985.

[10] H.W.Lenstra, C.Pomerance, and J.P.Buhler. *Factoring integers with the number field sieve.* In The Development of the Number Field Sieve, volume 1554, pages 50–94, New York, 1993. Lecture Notes in Mathematics, Springer-Verlag.

[11] A.K.Lenstra. *Integer factoring.* Designs, Codes and Cryptography, 19(2-3):101–128, 2000.

[12] L.T.Yang, L.Xu, M.Lin, J.Quinn. *A Parallel GNFS Algorithm with the Biorthogonal Block Lanczos Method for Integer Factorization.* Lecture Notes in Computer Science Volume 921, pp 106–120, 2006

[13] L.Xu, L.T.Yang, and M.Lin. *Parallel general number field sieve method for integer factorization.* In Proceedings of the 2005 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA-05), pages 1017–1023, Las Vegas, USA, June 2005.

[14] L.T.Yang, L.Xu, and M.Lin. *Integer factorization by a parallel gnfs algorithm for public key cryptosystem.* In Proceedings of the 2005 International Conference on Embedded Software and Systems (ICES-05), pages 683–695, Xian, China, December 2005.

[15] L.T.Yang, L.Xu, and S.Yeo and S.Hussain. *An integrated parallel {GNFS} algorithm for integer factorization based on Linbox Montgomery block Lanczos method over GF(2).* ScienceDirect, Computers & Mathematics with Applications, volume 60, number 2, pages 338 – 346, 2010.

[16] MPICH1: <http://www.mpich.org/>.

[17] T.Granlund. The GNU Multiple Precision Arithmetic Library (GNU MP). TMG Datakonsult, Boston, MA, USA, 2.0.2 edition, June 1996, http://www.ontko.com/pub/rayo/gmp/gmp_toc.html.

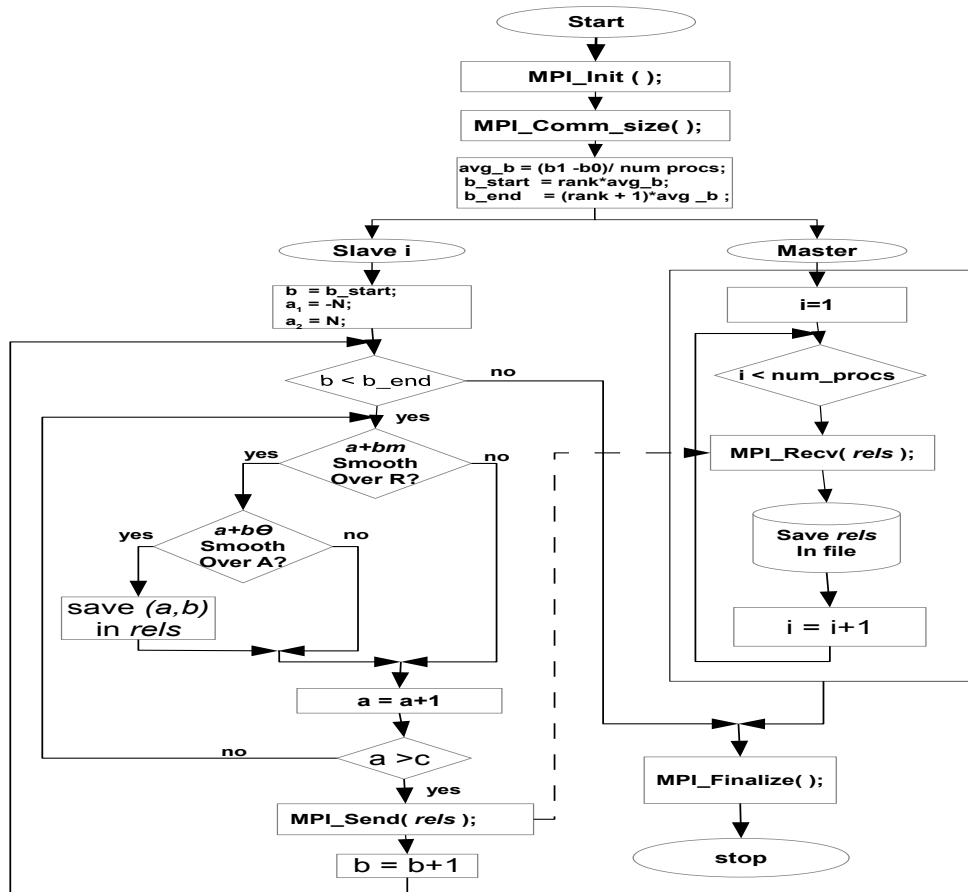


Fig. 2: The flowchart of second method

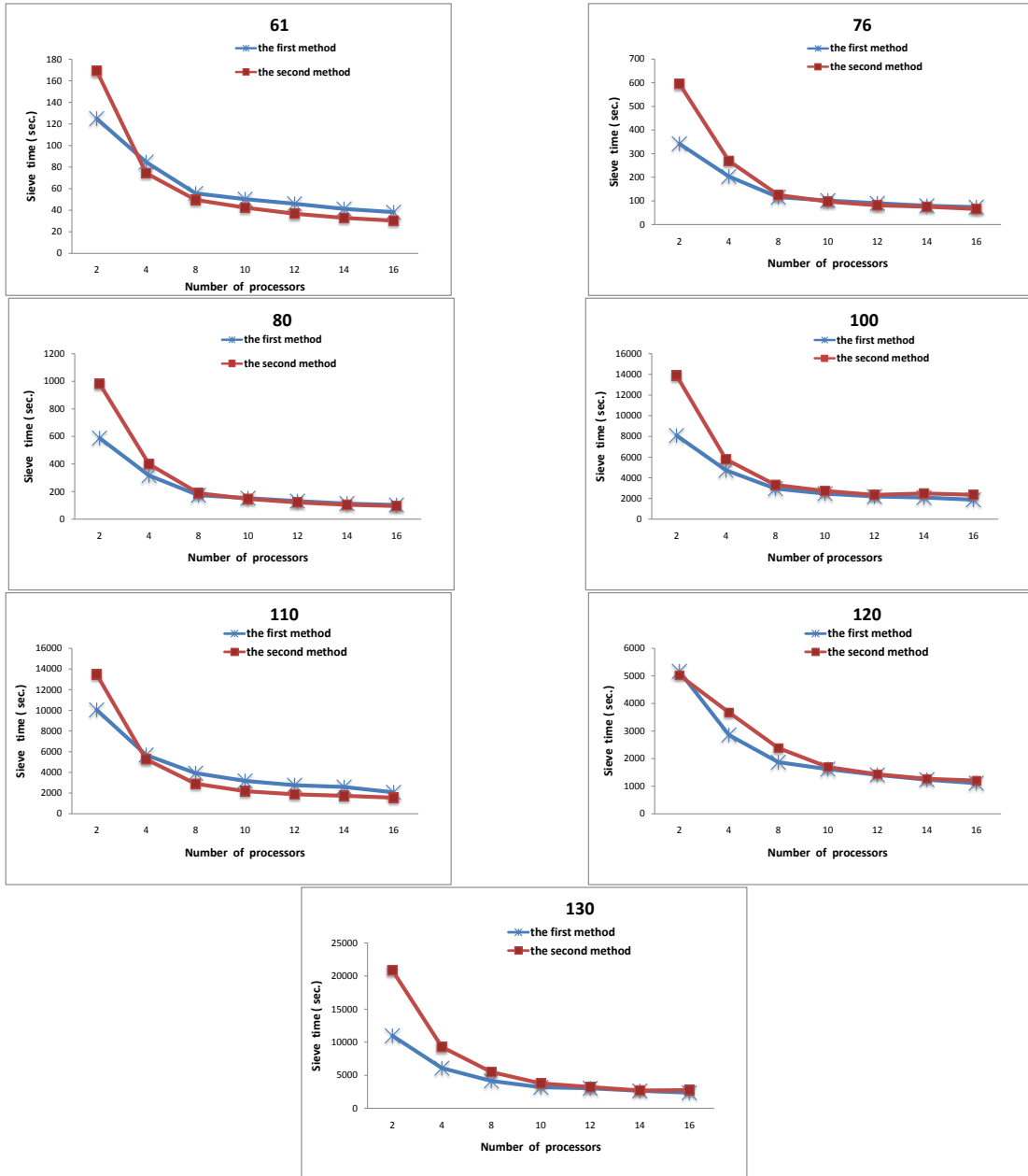


Fig. 3: The parallel implementation for the first and the second method. n = 61, 76, 80, 100, 110, 120, 130 digits

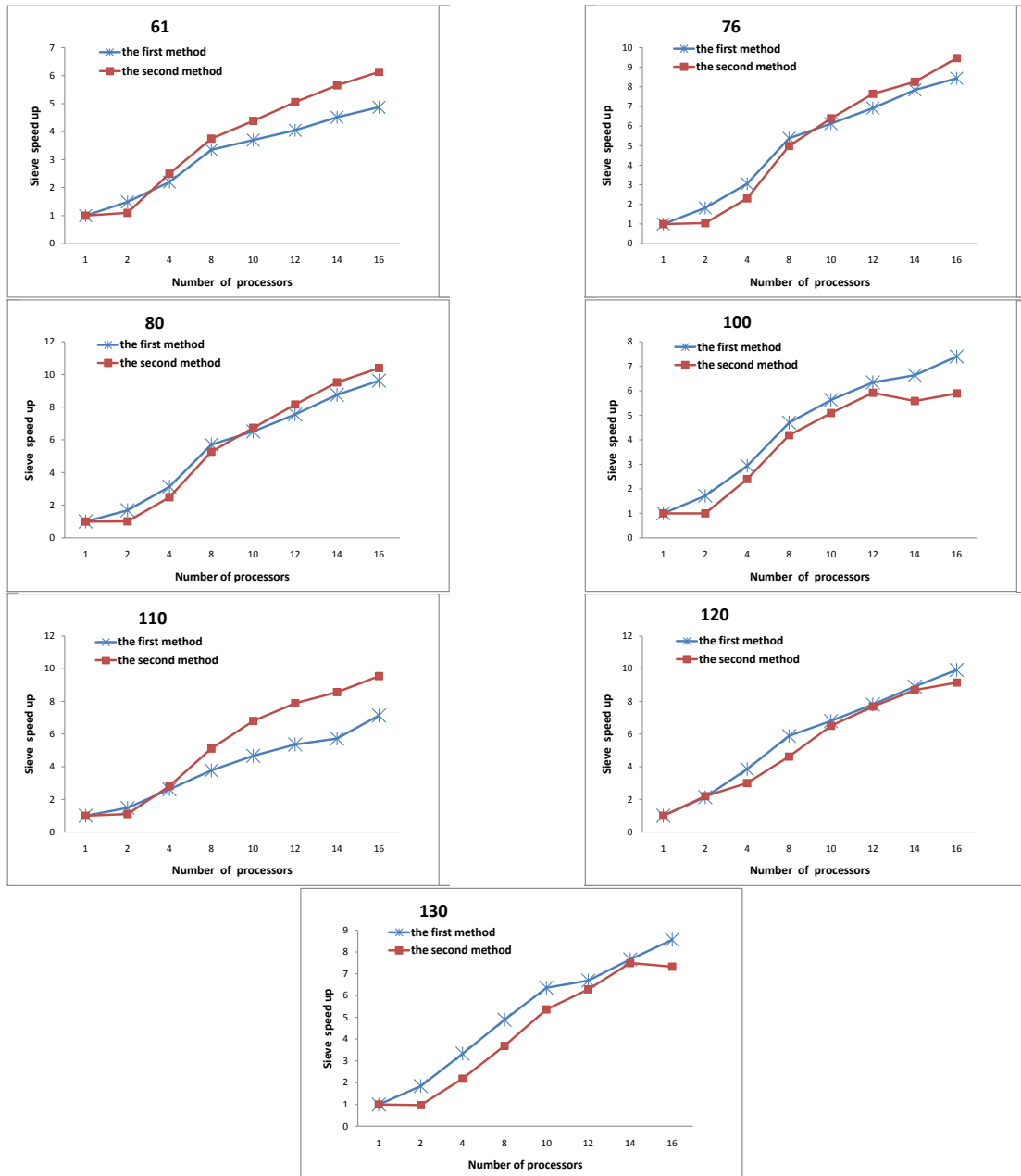


Fig. 4: Sieve Speed-up for the first method and the second method. n = 61, 76, 80, 100, 110, 120, 130 digits

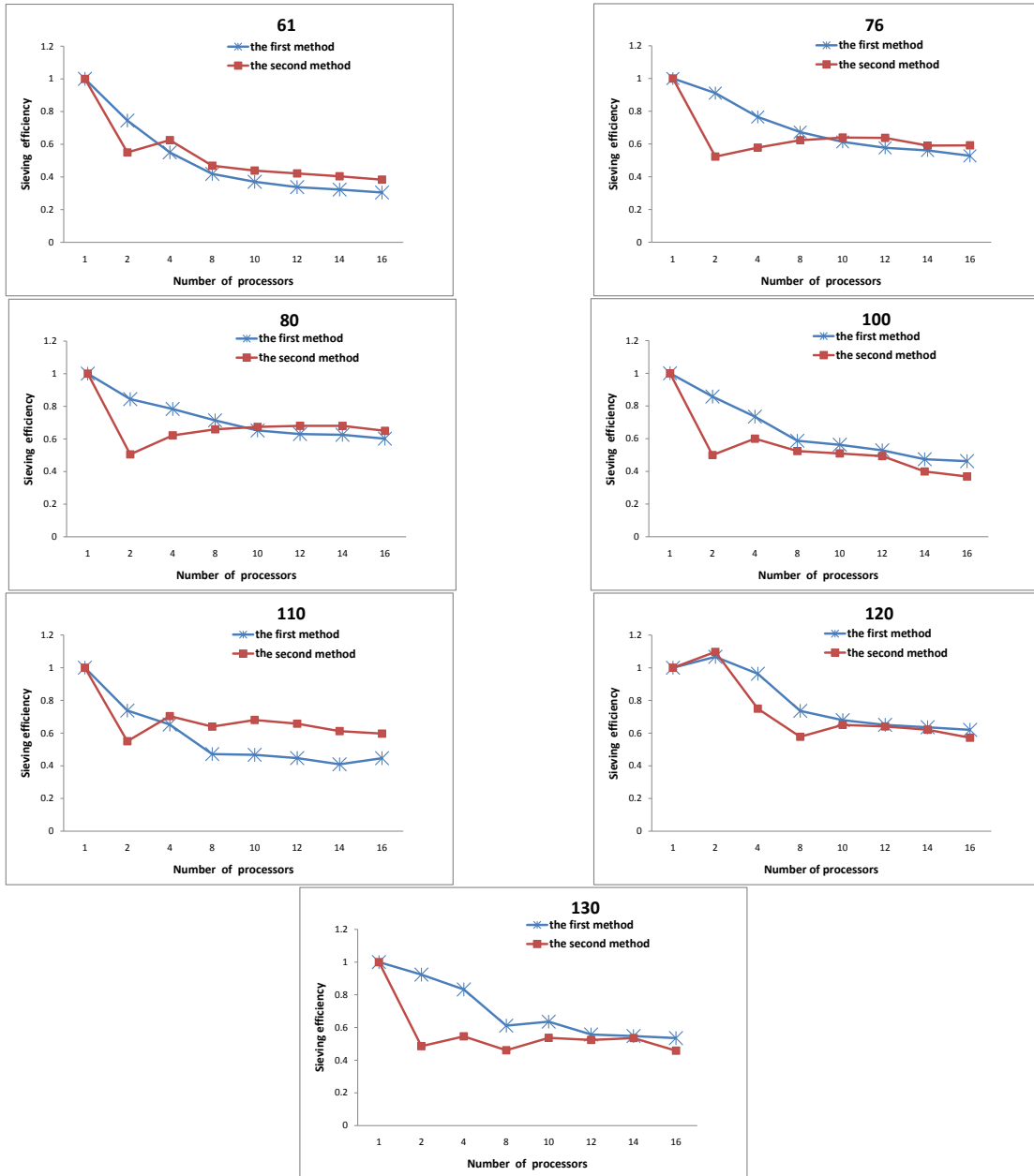


Fig. 5: Sieve efficiency for the first method and the second method. $n = 61, 76, 80, 100, 110, 120, 130$ digits

Estimating the Number of Test Workers Necessary for a Software Testing Process Using Artificial Neural Networks

Alaa F. Sheta

Faculty of Business and Technology
Princess Sumaya University for Technology
Amman, Jordan
asheta66@gmail.com

Sofian Kassaymeh

Information Technology Department
Taif University
Taif, Saudi Arabia
samsaak@gmail.com

David Rine

Emeritus of Computer Science
George Mason University
Fairfax, VA, USA
davidcrine@yahoo.com

Abstract—On time and within budget software project development represents a challenge for software project managers. Software management activities include but are not limited to: estimation of project cost, development of schedules and budgets, meeting user requirements and complying with standards. Recruiting development team members is a sophisticated problem for a software project manager. Since the utmost cost in software development effort is manpower, software project effort and is associated cost estimation models are used in estimating the effort required to complete a project. This effort estimate can then be converted into dollars based on the proper labor rates. An initial development team needs to be selected not only at the beginning of the project but also during the development process. It is important to allocate the necessary team to a project and efficiently distribute their effort during the development life cycle. In this paper, we provide our initial idea of developing a prediction model for defining the estimated required number of test workers of a software project during the software testing process. The developed models utilize the test instance and the number of observed faults as input to the proposed models. Artificial Neural Networks (ANNs) successfully build the dynamic relationships between the inputs and output and produce and accurate predication estimates.

Keywords—Staff Management; Neural Networks; Software Testing; Estimation

I. INTRODUCTION

"Software is a place where dreams are planted and nightmares harvested ... a world of were-wolves and silver bullets." This quote from Brad Cox [1] defined the challenges for software project managers in the past as well as today. The software project manager needs to have the skills, techniques and monitoring and control tools to meet the goal of a software development project. The goal is to complete software development within the agreed upon cost, schedule and user expectations. The measure of meeting this goal includes: meeting a schedule and a cost through improving budget distribution, managing human resources and adapting to environment changes. Intelligent project management requires many talents and skills. In 1987, the IEEE standards provide the following definition of software project management: Software project management is the

process of planning, organizing, staffing, monitoring, controlling, and leading a software project.

Software development has long been perceived as a risky business [2], [3]. A project manager can always try to predict the required resources and plan a schedule for a deliverable, but there is no guarantee that this is what will happen unless a careful monitoring and control plan is maintained. His/her ability to identify risks in advance could help planning for additional time to recover and reduce the consequence losses. According to Dr. Patricia Sanders, Director of Test Systems Engineering and Evaluation at OUSD, in her 1998 Software Technology Conference keynote address, 40% of the DoDs software development costs are spent on reworking the software, which in the year 2000 is equal to an actual loss of \$18 billion. Furthermore, Sanders stated that only 16% of software development would finish on time and on budget. It was also stated in [4] that:

Given that software-intensive projects are among the most expensive and risky undertakings of the 21st century, the investment in weapons from fiscal years 2003 through 2009 will exceed \$1 trillion. Furthermore, many of the DoD's most important technology projects will continue to deliver less than promised unless changes are made. Improving how we acquire software-intensive systems is both long overdue and an imperative.

In fact, the software development process is all about people, methodologies and tools. This can be seen from the software development process shown in Figure 1. People have to understand the project requirements, develop project plan and make a design, deployment of the project, test and validate the business requirements and finally fix bugs if any.

Software life cycle includes testing of the software system. The testing process requires significant effort and could cost over 50% of the project effort. This process requires a significant effort. It is defined as the process of executing a program with the intent of finding software bugs, errors or any defects [5]–[7]. It is also the process of validating and verifying that the developed software program will work and satisfies the needs of stakeholders. Software testing to be implemented needs a team of qualified personal. The team

Prof. A. Sheta is on leave from the Computers and Systems Department, Electronics Research Institute (ERI), Cairo, Egypt.

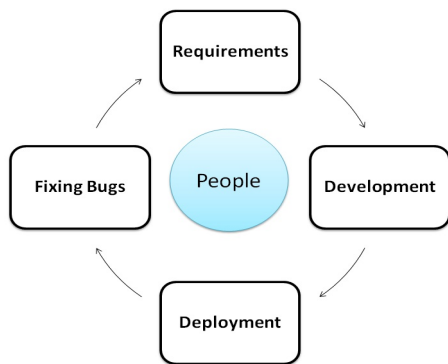


Fig. 1. Software development process

size depends on many factors. These factors include the size and complexity of the developed software or program. Staff turnover means frequent replacement of the development personnel. This in fact is one of the significant problems a software project manager could deal with.

In this paper, we provide a non-parametric Artificial Neural Network (ANN) model for predicting the number of test workers required during the software testing process. The number of required test workers will depend upon the count of faults (defects) observed at certain test instances. The model should be capable of accurately defining the required team size for testing and also help project managers distribute the effort of his team on various tasks required for the project. In Section II, we present a definition to the staff management problem. Statistical Regression Analysis is presented in Section III. An overview of soft computing techniques and specifically Artificial Neural Networks is presented in Section IV. The evaluation criterion for measuring the goodness of the developed models are presented in Section VI. The two case studies considered in this article are presented in Sections VII and VIII. Finally, we present the conclusion and future work.

II. STAFF MANAGEMENT

Time, cost, and number of staff estimations are essential duties for project managers in all business enterprises and especially for software projects. The manager needs to calculate an estimate for these main attributes in the early development process. This is not always an easy task for project managers. The role of a project manager is to manage, analyze and make decisions at all development phases according to accessible resources. Estimating time, cost, and staff helps sustain the monitoring and controlling of project activities and, in the end, produce quality. The field of software effort/cost estimation is concerned with providing an estimate of the expected cost, schedule, and manpower required to produce a software system. In fact there are common problems which could occur whenever we build a software system. The source of these problems could be one of the following:

- Insufficient requirements for the project
- Inadequate financial resources
- Loss of coordination because of many vendors

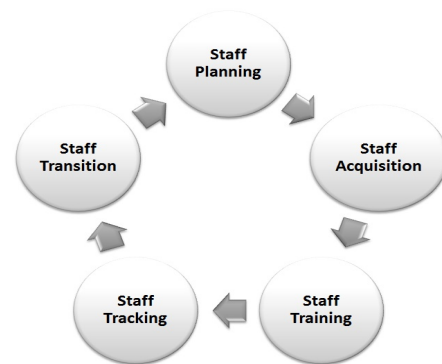


Fig. 2. Staff management process

- Staff turnover

The staff management process for the project consists of the following five elements: Staff Planning, Staff Acquisition, Staff Training, Staff Tracking, and Staff Transition. This process is shown in Figure 2. Specific information related to staff need to be collected, organized and updated during the project development life cycle. The staff management process for the project consists of the following five elements: Staff Planning, Staff Acquisition, Staff Training, Staff Tracking, and Staff Transition. This process is shown in Figure 2. Specific information related to staff needs is to be collected, organized and updated during the project development life cycle. The staff management information collected should include:

- The adequate numbers of staff needed for each project phase.
- The contribution as a function of time staff member.
- The source of staff such as staff hiring, part timer hiring or consulting.
- The schedule for joining and leaving the project.

Staff management includes project cost. The manager needs to gather adequate information such that the estimated project cost can be computed. Many software effort/cost estimation models were proposed to help in providing a high quality estimate to assist a project manager in considering the best decisions for a project [8], [9]. Many software cost estimation models were reported in the literature [10]–[13]. These models were used to help project managers to estimate effort, time and cost.

Staff scarcities is considered as sources for either inefficient use of resources or delay in delivering the project. Computing staffing members needed for a project depends on correct predictions of the project demand and expected date of the product to be in the market. Any delay might cause business loss or damage to firm reputation. Numerous methods were used to compute the estimate and predict staffing needs, based on the firms past experience, project types and sales and manufacture statistics [14]–[17].

III. STATISTICAL REGRESSION ANALYSIS

Statistical regression analysis associates relationships among a set of independent variables and one or more

dependent variables. The independent variables could be historical measurements about certain events in the past while we want to estimate or predict an independent variable at this instant of time or even in the future. Many techniques for carrying out regression analysis were evolved in the past. Linear regression and ordinary least squares regression are parametric methods that use Least Square Estimation (LSE) to estimate mathematical model parameters. COCOMO uses such regression methods.

A. Single Linear Regression

Regression analysis measures the degree of influence of the independent variables on a dependent variable. In the case of simple bivariate regression where there is a single independent variable, the dependent variable could be predicted from the independent variable by the simple equation:

$$y = a + bx + \epsilon \quad (1)$$

a is constant and b is the slope. This model is linear in the parameters a_i . y is called the independent variable and $x_i, i = 1, \dots, n$ are called the independent variables. The goal is to find the relationship between the dependent and independent variables. To compute the regression coefficient for the single independent variable given in Equation 1, we use the formula:

$$b = \frac{\sum(x_i - \hat{x})(y_i - \hat{y})}{\sum(x_i - \hat{x})^2} \quad (2)$$

Where \hat{x} is the mean (average) of the x values and \hat{y} is the mean of the y values. The parameter a is computed by the formula:

$$a = y - bx \quad (3)$$

Equation 2 can be expanded to be:

$$b = \frac{(\sum y_i \sum x_i^2) - (\sum x_i \sum x_i y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \quad (4)$$

B. Multiple Linear Regression

Equation 1 can be expanded to a multivariate concept as follows:

$$y = a_1 x_{i1} + a_2 x_{i2} + \dots + a_n x_{ij} \quad (5)$$

Where x_{ij} is the i^{th} observation on the j^{th} independent variable. To show how the parameter estimation process works, we assume we have a system with four input variables x_1, x_2, x_3, x_4 and single output y . Thus, the model mathematical equation can be represented as:

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 \quad (6)$$

To find the values of the model parameters a 's we need to build what is called the regression matrix ϕ . This matrix is developed based on the experiment collected measurements.

Thus, ϕ can be presented as follows given there is a set of measurements m :

$$X = \begin{pmatrix} 1 & x_1^1 & x_2^1 & x_3^1 & x_4^1 \\ 1 & x_1^2 & x_2^2 & x_3^2 & x_4^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^m & x_2^m & x_3^m & x_4^m \end{pmatrix}$$

The parameter vector θ and the output vector y can be presented as follows:

$$\theta = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} \quad y = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^m \end{pmatrix} \quad (7)$$

The least squares solution of yields the normal equation:

$$\phi^T \theta = y \quad (8)$$

which has a solution:

$$\theta = \phi^{-1} y \quad (9)$$

But since, the regression matrix ϕ is not a symmetric matrix, we have to reformulate the equation such that the solution for the parameter vector θ is as follows:

$$\theta = (\phi^T \phi)^{-1} \phi^T y \quad (10)$$

IV. SOFT-COMPUTING TECHNIQUES

Soft Computing techniques were explored to build efficient effort estimation models structures [18], [19]. In the past, authors in [20] explored the use of Neural Networks (ANNs), Genetic Algorithms (GAs) and Genetic Programming (GP) to provide a methodology for software cost estimation. ANN were used for software engineering project management in [21]. Authors in [22], provided a detailed study on using Genetic Programming (GP), Neural Network (ANNs) and Linear Regression (LR) in solving the software project estimation. Many data sets provided in [23], [24] were explored with promising results. A fuzzy COCOMO model was developed in [18].

Recently, In [12], author provided a pioneering set of models modified from the famous COCOMO model with interesting results. Later on, many authors explored the same idea with some modification [25]–[28] and provided a comparison to the work presented in [12]. Exploration of the advantages of the Takagi-Sugeno (TS) technique on building a set of linear models over the domain of possible software Kilo Line Of Code (KLOC) were investigated in [29]. Authors in [30] presented an extended work on the use of Soft Computing Techniques to build a suitable model structure to utilize improved estimations of software effort for NASA software projects. On doing this, Particle Swarm Optimization (PSO) was used to tune the parameters of the COCOMO model. The performance of the developed model was evaluated using NASA software projects data set. A comparison between COCOMO-PSO, Artificial Neural Networks (ANNs), Halstead, Walston-Felix, Bailey-Basili

and Doty models were provided with excellent modeling results. In [31], a research work describes the Estimation of Projects in Contexts of Uncertainty (EPCU) model. The model is an estimation process based on fuzzy logic which has the objective of solving the project estimation problem taking the benefits of the Expert Judgment in a formal way, without using quantitative historic data.

V. WHAT IS ANN?

According to the Defense Advanced Research Projects Agency (DARPA) Neural Network Study (1988, AFCEA International Press, p. 60):

... a neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes.

According to Nigrin (1993), p. 11 Nigrin1993, ANN was defined as:

A neural network is a circuit composed of a very large number of simple processing elements that are neurally based. Each element operates only on local information. Furthermore each element operates asynchronously; thus there is no overall system clock.

ANN can exhibit many brain-like behaviors such as learning, association, generalization, feature extraction, optimization and noise immunity. The basic simple unit of any ANN is the perceptron which is presented in Figure 3.

Artificial neural networks (ANN) have been proposed in many articles as a tool which was successfully able to develop software cost estimates. In [32], author provided a novel artificial neural network (ANN) prediction model which incorporates COCOMO and ANN-COCOMO II, to provide more accurate software estimates at the early phase of software development. ANN was employed to regulate the software features considering historical project data. In [33], authors provided a survey on the cost estimation models using artificial neural networks. ANN has many advantages they include:

- A neural network can perform tasks that a linear program cannot.
- When an element of the neural network fails, it can continue without any problem by their parallel nature.
- A neural network learns and does not need to be reprogrammed.
- It can be implemented in any application.

The learning process in ANN is the algorithm which is used to adjust the weights of the network in order to minimize the difference between the actual and predicted values by the network. Usually, the weights of the network are initialized randomly. There are four basic types of learning rule: Error Correlation Learning (ECL), Boltzmann

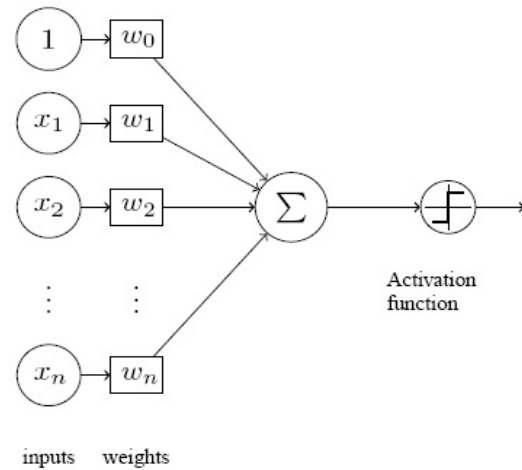


Fig. 3. The simple building block of ANN

learning (BL), Hebbian Learning (HL), and Competitive Learning (CL). The detailed descriptions of these learning rules are referred to the work of [34]. Among all the training algorithms, Back-Propagation (BP) which follows ECL rule is the most popular choice.

VI. EVALUATION CRITERIA

The performance of the developed two models; the Auto-Regression and the Artificial Neural Networks models will be evaluated using a number of evaluation criteria. They are:

- The Variance-Accounted-For (VAF) criteria was adopted by [35]:

$$VAF = \left[1 - \frac{\text{var}(y - \hat{y})}{\text{var}(y)} \right] \times 100\% \quad (11)$$

- The Mean square error (MSE):

$$MSE = \frac{1}{N} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (12)$$

- The Euclidian distance (ED):

$$ED = \sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (13)$$

- The Manhattan distance (MD):

$$MD = \frac{1}{N} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (14)$$

- In [36], the authors provided an empirical study for data modeling in software engineering application and used radial basis function (RBF) to develop effort estimation model. They considered the mean magnitude of relative error (MMRE) as the main performance measure. We will evaluate the (MMRE) over the training and testing data as described in [36]. The mean magnitude of relative error (MMRE), defined as:

TABLE I. TEST/DEBUG DATA 1 x_1 : TEST INSTANCES, x_2 : REAL DETECTED FAULTS, y_k : NO. OF TEST WORKERS

| x_1 | x_2 | y | x_1 | x_2 | y | x_1 | x_2 | y |
|-------|-------|-----|-------|-------|-----|-------|-------|-----|
| 1 | 5 | 4 | 38 | 15 | 8 | 75 | 0 | 4 |
| 2 | 5 | 4 | 39 | 7 | 8 | 76 | 0 | 4 |
| 3 | 5 | 4 | 40 | 15 | 8 | 77 | 1 | 4 |
| 4 | 5 | 4 | 41 | 21 | 8 | 78 | 2 | 2 |
| 5 | 6 | 4 | 42 | 8 | 8 | 79 | 0 | 2 |
| 6 | 8 | 5 | 43 | 6 | 8 | 80 | 1 | 2 |
| 7 | 2 | 5 | 44 | 20 | 8 | 81 | 0 | 2 |
| 8 | 7 | 5 | 45 | 10 | 8 | 82 | 0 | 2 |
| 9 | 4 | 5 | 46 | 3 | 8 | 83 | 0 | 2 |
| 10 | 2 | 5 | 47 | 3 | 8 | 84 | 0 | 2 |
| 11 | 31 | 5 | 48 | 8 | 4 | 85 | 0 | 2 |
| 12 | 4 | 5 | 49 | 5 | 4 | 86 | 0 | 2 |
| 13 | 24 | 5 | 50 | 1 | 4 | 87 | 2 | 2 |
| 14 | 49 | 5 | 51 | 2 | 4 | 88 | 0 | 2 |
| 15 | 14 | 5 | 52 | 2 | 4 | 89 | 0 | 2 |
| 16 | 12 | 5 | 53 | 2 | 4 | 90 | 0 | 2 |
| 17 | 8 | 5 | 54 | 7 | 4 | 91 | 0 | 2 |
| 18 | 9 | 5 | 55 | 2 | 4 | 92 | 0 | 2 |
| 19 | 4 | 5 | 56 | 0 | 4 | 93 | 0 | 2 |
| 20 | 7 | 5 | 57 | 2 | 4 | 94 | 0 | 2 |
| 21 | 6 | 5 | 58 | 3 | 4 | 95 | 0 | 2 |
| 22 | 9 | 5 | 59 | 2 | 4 | 96 | 1 | 2 |
| 23 | 4 | 5 | 60 | 7 | 4 | 97 | 0 | 2 |
| 24 | 4 | 5 | 61 | 3 | 4 | 98 | 0 | 2 |
| 25 | 2 | 5 | 62 | 0 | 4 | 99 | 0 | 2 |
| 26 | 4 | 5 | 63 | 1 | 4 | 100 | 1 | 2 |
| 27 | 3 | 5 | 64 | 0 | 4 | 101 | 0 | 1 |
| 28 | 9 | 6 | 65 | 1 | 4 | 102 | 0 | 1 |
| 29 | 2 | 6 | 66 | 0 | 3 | 103 | 1 | 1 |
| 30 | 5 | 6 | 67 | 0 | 3 | 104 | 0 | 1 |
| 31 | 4 | 6 | 68 | 1 | 3 | 105 | 0 | 1 |
| 32 | 1 | 6 | 69 | 1 | 3 | 106 | 1 | 1 |
| 33 | 4 | 6 | 70 | 0 | 3 | 107 | 0 | 1 |
| 34 | 3 | 6 | 71 | 0 | 3 | 108 | 0 | 1 |
| 35 | 6 | 6 | 72 | 1 | 3 | 109 | 1 | 1 |
| 36 | 13 | 6 | 73 | 1 | 4 | 110 | 0 | 1 |
| 37 | 19 | 8 | 74 | 0 | 4 | 111 | 1 | 1 |

TABLE II. EVALUATION CRITERIA OF THE ANN MODELS

| Criteria | VAF | MSE | ED | MD | MMRE |
|----------|---------|-------|-------|-------|-------|
| MR | 85.621% | 0.559 | 7.881 | 0.526 | 0.148 |
| NN | 96.347% | 0.143 | 3.994 | 0.243 | 0.076 |

$$MMRE = \frac{1}{N} \sum_{i=1}^N \frac{|y_i - \hat{y}_i|}{y_i} \quad (15)$$

Where y and \hat{y} are the observed and predicted number of test workers the neural network model and n is the number of measurements used in the experiments, respectively.

VII. TEST/DEBUG DATA 1

Field report data was developed to measure system faults during testing in a real-time application [37]. The software system consists of 200 modules with each having one kilo line of code of FORTRAN. A Test/Debug dataset of 111 measurements is given in Table I. To develop a ANN test work estimate model, we used the data set to train the ANN. The observed and predicted number of workers was calculated based on the test instances and the real detected faults and shown in Figure 4. The convergence of the neural networks is shown in Figure 5 over 3000 epochs. The observed and predicted number of workers calculated based the test instances and the real detected faults is shown in Figure 5. The convergence of the neural networks is shown in Figure 6.

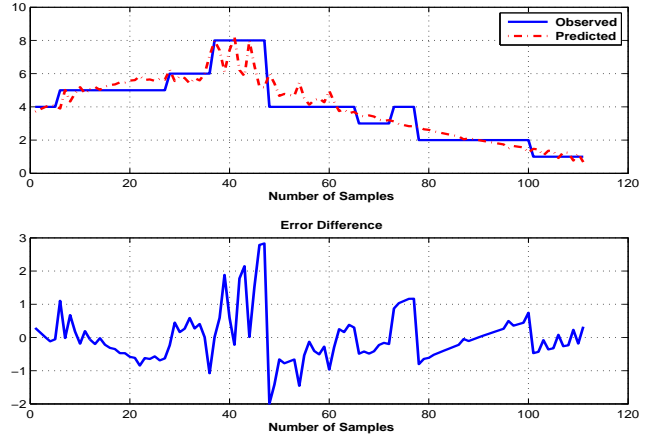


Fig. 4. Observed and predicted number of test workers using Multiple Regression Model: Test/Debug Data 1

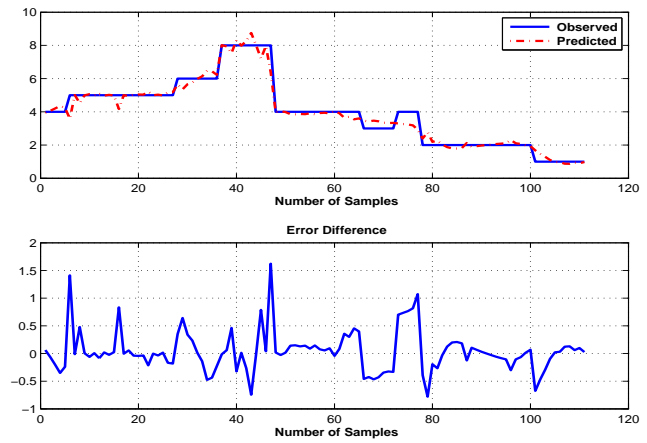


Fig. 5. Observed and predicted number of test workers using ANN: Test/Debug Data 1

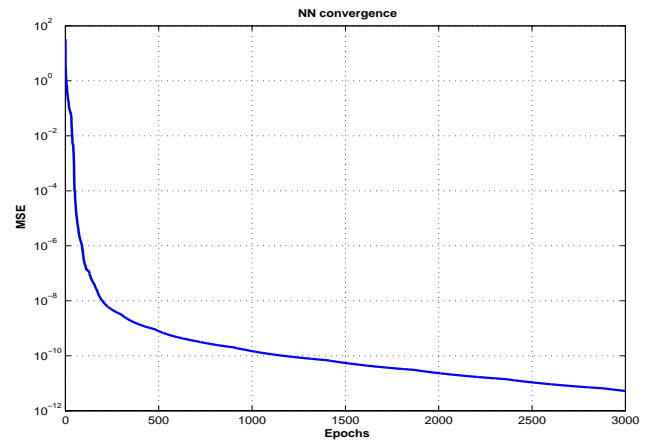


Fig. 6. NN Convergence using 50 neurons in the hidden layer: Test/Debug Data 1

TABLE III. TEST/DEBUG DATA 2 x_1 : TEST INSTANCES, x_2 : REAL DETECTED FAULTS, y_k : NO. OF TEST WORKERS

| x_1 | x_2 | y | x_1 | x_2 | y |
|-------|-------|-----|-------|-------|-----|
| 1 | 2 | 75 | 24 | 2 | 8 |
| 2 | 0 | 31 | 25 | 1 | 15 |
| 3 | 30 | 63 | 26 | 7 | 31 |
| 4 | 13 | 128 | 27 | 0 | 1 |
| 5 | 13 | 122 | 28 | 22 | 57 |
| 6 | 3 | 27 | 29 | 2 | 27 |
| 7 | 17 | 136 | 30 | 5 | 35 |
| 8 | 2 | 49 | 31 | 12 | 26 |
| 9 | 2 | 26 | 32 | 14 | 36 |
| 10 | 20 | 102 | 33 | 5 | 28 |
| 11 | 13 | 53 | 34 | 2 | 22 |
| 12 | 3 | 26 | 35 | 0 | 4 |
| 13 | 3 | 78 | 36 | 7 | 8 |
| 14 | 4 | 48 | 37 | 3 | 5 |
| 15 | 4 | 75 | 38 | 0 | 27 |
| 16 | 0 | 14 | 39 | 0 | 6 |
| 17 | 0 | 4 | 40 | 0 | 6 |
| 18 | 0 | 14 | 41 | 0 | 4 |
| 19 | 0 | 22 | 42 | 5 | 1 |
| 20 | 0 | 5 | 43 | 2 | 6 |
| 21 | 0 | 9 | 44 | 3 | 5 |
| 22 | 30 | 33 | 45 | 0 | 8 |
| 23 | 15 | 18 | 46 | 0 | 2 |

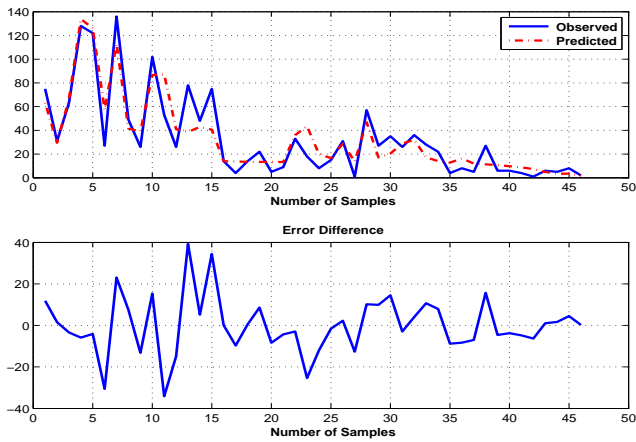


Fig. 7. Observed and predicted number of test workers using Multiple Regression Model: Test/Debug Data 2

VIII. TEST/DEBUG DATA 2

A Test/Debug data set has 46 measurements is given in Table III. The data set was presented in [37]. The number of measurements collected during the testing process is small. This represents a difficulty for traditional parameter estimation techniques. It is sometimes difficult to correctly estimate model parameters using a small number of measurements. To build a test work estimate model, we used the data set to build both the MR and ANN models. The observed and predicted number of workers calculated is based on the test instances and the real detected faults are shown in Figure 8. The convergence of the neural networks is shown in Figure 9 over 3000 epochs.

TABLE IV. EVALUATION CRITERIA OF THE ANN MODELS

| Criteria | VAF | MSE | ED | MD | MMRE |
|----------|---------|--------|--------|-------|-------|
| MR | 84.088% | 188.73 | 93.175 | 9.992 | 0.931 |
| NN | 89.098% | 129.99 | 77.328 | 7.252 | 0.707 |

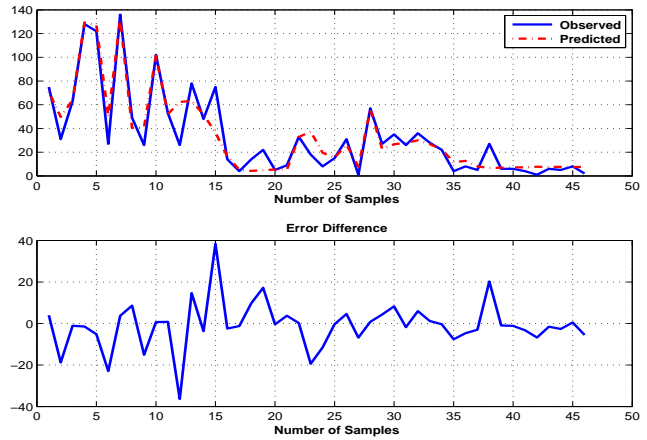


Fig. 8. Observed and predicted number of test workers using ANN: Test/Debug Data 2

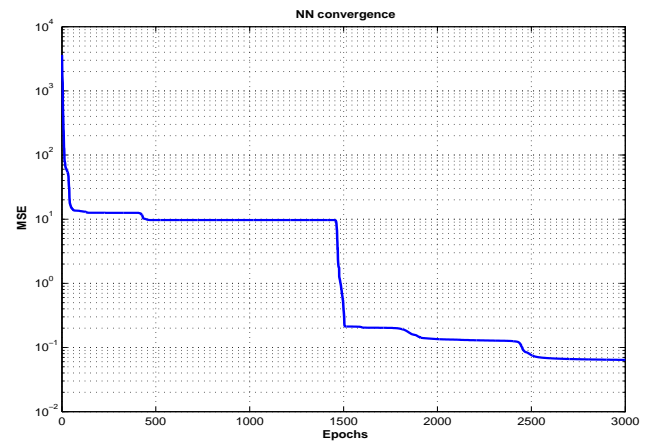


Fig. 9. NN convergence using 20 neurons in the hidden layer: Test/Debug Data 2

IX. CONCLUSIONS AND FUTURE WORK

Estimating the number of test workers during the software testing process became a challenge problem. Numerous methods were used to estimate and predict staffing needs, based on the firms past experience, project types and sales and manufacture statistics. Thus, tools and methods are required to fill the gap in this major area of software project life cycle development. In this paper, we propose our initial idea of developing predictive models for defining the estimated number of test workers of a software project during the software testing process using ANN. The developed models utilize the test instance and the number of observed faults as input to the proposed models. Two cases studies were presented and many evaluation criteria were used to validate the developed model performance. Artificial Neural Networks (ANNs) successfully build the dynamic relationships between the inputs and output and produce and accurate predication estimates. We plan to explore other soft computing techniques to handle this problem such as fuzzy logic to develop a mathematical relationship which can be easily explained in this case.

REFERENCES

- [1] B. J. Cox, "Planning the software industrial revolution," vol. 7, no. 6. IEEE Software, 1990, pp. 25–33.
- [2] W.-M. Han and S.-J. Huang, "An empirical analysis of risk components and performance on software projects," *J. Syst. Softw.*, vol. 80, no. 1, pp. 42–50, Jan. 2007.
- [3] N. Ramasubbu and R. K. Balan, "Overcoming the challenges in cost estimation for distributed software projects," in *Proceedings of the 34th International Conference on Software Engineering*, ser. ICSE '12. Piscataway, NJ, USA: IEEE Press, 2012, pp. 91–101. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2337223.2337235>
- [4] L. Pracchia, "Improving the dod: Software acquisition processes," *The Journal of Defense Software Engineering*, vol. 4, pp. 4–7, 2004.
- [5] M. B. Cohen, M. B. Dwyer, and J. Shi, "Coverage and adequacy in software product line testing," in *Proceedings of the ISSSTA 2006 Workshop on Role of Software Architecture for Testing and Analysis*, ser. ROSATEA '06. New York, NY, USA: ACM, 2006, pp. 53–63. [Online]. Available: <http://doi.acm.org/10.1145/1147249.1147257>
- [6] I.-C. Yoon, A. Sussman, A. Memon, and A. Porter, "Effective and scalable software compatibility testing," in *Proceedings of the 2008 International Symposium on Software Testing and Analysis*, ser. ISSTA '08. New York, NY, USA: ACM, 2008, pp. 63–74. [Online]. Available: <http://doi.acm.org/10.1145/1390630.1390640>
- [7] T. Bultan and T. Xie, "Workshop on testing, analysis and verification of web software (tav-web 2008)," in *Proceedings of the 2008 International Symposium on Software Testing and Analysis*, ser. ISSTA '08. New York, NY, USA: ACM, 2008, pp. 311–312. [Online]. Available: <http://doi.acm.org/10.1145/1390630.1390670>
- [8] B. Boehm, *Software Engineering Economics*. Englewood Cliffs, NJ, Prentice-Hall, 1981.
- [9] —, *Cost Models for Future Software Life Cycle Process: COCOMO2*. Annals of Software Engineering, 1995.
- [10] K. Pillai and S. Nair, "A model for software development effort and cost estimation," *IEEE Trans. on Software Engineering*, vol. 23, pp. 485–497, 1997.
- [11] C. Schofield, "An empirical investigation into software effort estimation by analogy," Ph.D. dissertation, Bournemouth University, 1998.
- [12] A. F. Sheta, "Estimation of the COCOMO model parameters using genetic algorithms for NASA software projects," *Journal of Computer Science*, vol. 2, no. 2, pp. 118–123, 2006.
- [13] A. F. Sheta, A. Ayesh, and D. Rine, "Evaluating software cost estimation models using particle swarm optimisation and fuzzy logic for nasa projects: a comparative study," *Int. J. Bio-Inspired Comput.*, vol. 2, no. 6, pp. 365–373, Nov. 2010. [Online]. Available: <http://dx.doi.org/10.1504/IJBIC.2010.037016>
- [14] T. P. Bechet, *Strategic Staffing: A Comprehensive System for Effective Workforce Planning*. AMACOM Div American Mgmt Assn, 2008.
- [15] H. Zeng and D. Rine, "A neural network approach for software defects fix effort estimation," in *IASTED Conf. on Software Engineering and Applications*, M. H. Hamza, Ed. IASTED/ACTA Press, 2004, pp. 513–517.
- [16] U. Saxena and S. P. Singh, "Software effort estimation using neuro-fuzzy approach," in *2012 CSI Sixth International Conference on Software Engineering (CONSEG)*, 2012, pp. 1–6.
- [17] H. Zhang, L. Gong, and S. Versteeg, "Predicting bug-fixing time: An empirical study of commercial software projects," in *Proceedings of the 2013 International Conference on Software Engineering*, ser. ICSE '13. Piscataway, NJ, USA: IEEE Press, 2013, pp. 1042–1051. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2486788.2486931>
- [18] J. Ryder, "Fuzzy COCOMO: Software cost estimation," Ph.D. dissertation, Binghamton University, 1995.
- [19] A. C. Hodgkinson and P. W. Garratt, "A neuro-fuzzy cost estimator," in *Proceedings of the Third Conference on Software Engineering and Applications*, 1999, pp. 401–406.
- [20] M. A. Kelly, "A methodology for software cost estimation using machine learning techniques," Master's thesis, Naval Postgraduate School, Monterey, California, 1993.
- [21] S. Kumar, B. A. Krishna, and P. Satsangi, "Fuzzy systems and neural networks in software engineering project management," *Journal of Applied Intelligence*, vol. 4, pp. 31–52, 1994.
- [22] J. J. Dolado and L. F. andez, "Genetic programming, neural network and linear regression in software project estimation," in *Proceedings of the INSPIRE III, Process Improvement through training and education*. British Company Society, 1998, pp. 157–171.
- [23] A. J. Albrecht and J. R. Gaffney, "Software function, source lines of code, and development effort prediction: A software science validation," *IEEE Trans. Software Engineering*, vol. 9, no. 6, pp. 630–648, 1983.
- [24] J. E. Matson, B. E. Barret, and J. M. Mellinchamp, "Software development cost estimation using function points," *IEEE Trans. Software Engineering*, vol. 20, no. 4, pp. 275–287, 1994.
- [25] H. Mittal and P. Bhatia, "A comparative study of conventional effort estimation and fuzzy effort estimation based on triangular fuzzy numbers," *International Journal of Computer Science and Security*, vol. 1, no. 4, pp. 36–47, 2007.
- [26] —, "Optimization criteria for effort estimation using fuzzy technique," *CLEI Electronic Journal*, vol. 10, no. 1, pp. 1–11, 2007.
- [27] M. Uysal, "Estimation of the effort component of the software projects using simulated annealing algorithm," in *World Academy of Science, Engineering and Technology*, vol. 41, 2008, pp. 258–261.
- [28] P. S. Sandhu, M. Prashar, P. Bassi, and A. Bisht, "A model for estimation of efforts in development of software systems," in *World Academy of Science, Engineering and Technology*, vol. 56, 2009, pp. 148–152.
- [29] A. Sheta, "Software effort estimation and stock market prediction using takagi-sugeno fuzzy models," in *Proceedings of the 2006 IEEE Fuzzy Logic Conference, Sheraton, Vancouver Wall Centre, Vancouver, BC, Canada, July 16-21, 2006*, pp. 579–586.
- [30] A. Sheta, D. Rine, and A. Ayesh, "Development of software effort and schedule estimation models using soft computing techniques," in *Proceedings of the 2008 IEEE Congress on Evolutionary Computation (IEEE CEC 2008) within the 2008 IEEE World Congress on Computational Intelligence (WCCI 2008), Hong Kong, 1-6 June, 2008*, pp. 1283–1289.
- [31] F. V. Souto, *Design Of A Fuzzy Logic Estimation Process For Software Projects: Estimation of Projects in a Context of Uncertainty EPCU Model*. Germany: LAP Lambert Academic Publishing, 2012.
- [32] I. Attarzadeh, A. Mehrzadeh, and A. Barati, "Proposing an enhanced artificial neural network prediction model to improve the accuracy in software effort estimation," in *Proceedings of the 2012 Fourth International Conference on Computational Intelligence, Communication Systems and Networks*, ser. CICSYN '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 167–172.
- [33] M. Shepper and C. Schofield, "Estimating software project effort using analogies," *IEEE Tran. Software Engineering*, vol. 23, pp. 736–743, 1997.
- [34] A. K. Jain, J. Mao, and K. K. Mohiuddin, "Artificial neural networks: A tutorial," *IEEE Computer Special Issue on Neural Computing*, pp. 31–44, 1996.
- [35] R. Babuška, *Fuzzy Modeling and Identification Toolbox*. Delft University of Technology, The Netherlands, <http://lcewww.et.tudelft.nl/babuska>, 1998.
- [36] M. Shin and A. L. Goel, "Empirical data modeling in software engineering using radial basis functions," *IEEE Trans. on Software Engineering*, pp. 567–576, 2000.
- [37] Y. Tohman, K. Tokunaga, S. Nagase, and M. Y., "Structural approach to the estimation of the number of residual software faults based on the hyper-geometric distribution model," *IEEE Trans. on Software Engineering*, pp. 345–355, 1989.

Natural Gradient Descent for Training Stochastic Complex-Valued Neural Networks

Tohru Nitta

National Institute of Advanced Industrial Science and Technology (AIST),
AIST Tsukuba Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
Email: tohru-nitta@aist.go.jp

Abstract—In this paper, the natural gradient descent method for the multilayer stochastic complex-valued neural networks is considered, and the natural gradient is given for a single stochastic complex-valued neuron as an example. Since the space of the learnable parameters of stochastic complex-valued neural networks is not the Euclidean space but a curved manifold, the complex-valued natural gradient method is expected to exhibit excellent learning performance.

Keywords—Neural network; Complex number; Learning; Singular point

I. INTRODUCTION

Complex-valued neural networks whose parameters (weights and threshold values) are all complex numbers, are useful in fields dealing with complex numbers or two-dimensional vectors such as telecommunications, speech recognition and image processing with Fourier transformation. Indeed, we can find some applications of complex-valued neural networks to various fields in the literature [6], [9].

The multilayer complex-valued neural network is usually trained using the gradient descent learning method [5], [10], [11], [12], as in the case of the multilayer real-valued neural network. The space of the learnable parameters of stochastic complex-valued neural networks is, however, not the Euclidean space but a curved manifold. For stochastic complex-valued neural networks, the ordinary gradient does not give the steepest direction of a target function, and the steepest direction is given by the natural gradient [2], [3]. It has been shown in [4] that the natural gradient method could avoid singular points of the real-valued parameter space which is a cause of standstill in learning, and the natural gradient method could improve the learning performance of the real-valued neural networks as a result. Similarly, there exist many singular points in the complex-valued neural networks [7]. Thus, the natural gradient method would be useful for the complex-valued neural networks, too. In this paper, we extend the natural gradient descent method for the multilayer stochastic real-valued neural networks to the complex domain, and give the natural gradient for a single stochastic complex-valued neuron as an example.

Section II describes the complex-valued neural network. Section III is devoted to the explanation of the natural gradient method, and Section IV presents the natural gradient method in complex-valued neural networks, which is followed by our conclusion in Section V.

II. COMPLEX-VALUED NEURAL NETWORK MODEL

This section describes the complex-valued neural network model used in this paper. First, we will consider the following complex-valued neuron. The input signals, weights, thresholds and output signals are all complex numbers. The net input U_n to a complex-valued neuron n is defined as: $U_n = \sum_m W_{nm} X_m + V_n$, where W_{nm} is the complex-valued weight connecting the complex-valued neurons n and m , X_m is the complex-valued input signal from the complex-valued neuron m , and V_n is the complex-valued threshold value of the complex-valued neuron n . To obtain the complex-valued output signal, convert the net input U_n into its real and imaginary parts as follows: $U_n = x + iy = z$, where i denotes $\sqrt{-1}$. The complex-valued output signal is defined to be

$$f_C(z) = \varphi(x) + i\varphi(y), \quad (1)$$

where $\varphi: \mathbf{R} \rightarrow \mathbf{R}$, (\mathbf{R} denotes the set of real numbers). Eq. (1) is often called a *split-type* complex-valued activation function. Note that the activation function f_C is not a regular complex-valued function because the Cauchy-Riemann equations do not hold.

The complex-valued neural network used in this paper consists of such complex-valued neurons described above.

Note that various types of activation functions other than Eq. (1) can be considered naturally (for examples, the non-split-type (fully) one [10]).

III. NATURAL GRADIENT METHOD

This section briefly describes the natural gradient proposed in [2], [3]. Let $S = \{\mathbf{w} \in \mathbf{R}^N\}$ be a Riemannian space with the Riemannian metric tensor $\mathbf{G}(\mathbf{w}) = (g_{ij}(\mathbf{w}))$ on which a function $L(\mathbf{w})$ is defined. If

$$g_{ij}(\mathbf{w}) = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}, \quad (2)$$

that is, $\mathbf{G}(\mathbf{w})$ is the unit matrix, then S is an Euclidean space. Amari proved the following theorem [3].

Theorem 1: *The steepest descent direction of $L(\mathbf{w})$ in a Riemannian space is given by*

$$-\tilde{\nabla}L(\mathbf{w}) = -\mathbf{G}^{-1}(\mathbf{w})\nabla L(\mathbf{w}) \quad (3)$$

where $\mathbf{G}^{-1}(\mathbf{w}) = (g^{ij}(\mathbf{w}))$ is the inverse of the metric $\mathbf{G}(\mathbf{w}) = (g_{ij}(\mathbf{w}))$ and ∇L is the conventional gradient,

$$\nabla L(\mathbf{w}) = \left(\frac{\partial}{\partial w_1} L(\mathbf{w}), \dots, \frac{\partial}{\partial w_N} L(\mathbf{w}) \right)^T, \quad (4)$$

where the superscript T denotes the transposition. \square

$\tilde{\nabla}L(\mathbf{w}) = \mathbf{G}^{-1}\nabla L(\mathbf{w})$ is called the *natural gradient* of L in the Riemannian space. The natural gradient descent algorithm is given by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \varepsilon_t \tilde{\nabla}L(\mathbf{w}_t), \quad (5)$$

where ε_t is the learning rate.

Amari derived the natural gradients explicitly in the case of the space of real-valued perceptrons for neural learning, the space of matrices for blind source separation, and the space of linear dynamical systems for blind multichannel source deconvolution [3].

IV. NATURAL GRADIENT IN COMPLEX-VALUED NEURAL NETWORKS

In this section, the natural gradient is applied to the complex-valued neural networks and the natural gradient descent algorithm is explicitly derived for a single complex-valued neuron.

A. Natural Gradient Learning in Complex-Valued Neural Networks

Let us consider a stochastic complex-valued multilayer feedforward neural network with N input neurons, one output neuron, and a learnable complex-valued vector parameter $\mathbf{w} = (w_1, \dots, w_N)^T \in \mathbf{C}^N$ which consists of all the weights and thresholds (\mathbf{C} denotes the set of complex numbers). Assume that the complex-valued input signal $\mathbf{z} = (z_1, \dots, z_N)^T \in \mathbf{C}^N$ is subject to an unknown probability distribution $q(\mathbf{z})$, and the complex-valued output signal $y \in \mathbf{C}$ is given by

$$y = g_C(\mathbf{z}, \mathbf{w}) + n, \quad (6)$$

where g_C is a complex function, and $n = n_R + in_I$ is a complex-valued random variable subject to a complex normal distribution (or bivariate normal distribution) $N(\mu, \Sigma)$. The model specifies the probability density of the input-output pair as

$$p(\mathbf{z}, y; \mathbf{w}) = q(\mathbf{z}) \cdot p(y|\mathbf{z}; \mathbf{w}). \quad (7)$$

Define a loss function $l(\mathbf{z}, y; \mathbf{w})$ when input signal \mathbf{z} is processed by the stochastic complex-valued neural network having parameter \mathbf{w} as:

$$\begin{aligned} l(\mathbf{z}, y; \mathbf{w}) &\stackrel{\text{def}}{=} -\log p(\mathbf{z}, y; \mathbf{w}) \\ &= -\log q(\mathbf{z}) - \log p(y|\mathbf{z}; \mathbf{w}). \end{aligned} \quad (8)$$

Given the training set $\{(z_t, y_t), t = 1, \dots, T\}$, minimizing the loss function (Eq. (8)) is equivalent to maximizing the probability that the stochastic complex-valued neural network outputs the training output signal y_t .

The space of all the probability distributions which the above stochastic complex-valued neural network realizes, can be regarded as a $2N$ -dimensional Riemannian space because the complex-valued parameter consists of the two real-valued parameters: the real-part and the imaginary part. Thus, the information geometry [1] can be applied to the complex-valued case, too.

The natural gradient descent algorithm for the complex-valued neural network is given by

$$\mathbf{v}_{t+1} = \mathbf{v}_t - \varepsilon_t \tilde{\nabla}l(\mathbf{z}_t, y_t; \mathbf{v}_t), \quad (9)$$

where $\{(z_t, y_t) \in \mathbf{C}^N \times \mathbf{C}, t = 1, 2, \dots\}$ is the sequence of the complex-valued training signals, and

$$\begin{aligned} \mathbf{v} &= (v_1, \dots, v_{2N})^T \\ &= (\text{Re}[w_1], \dots, \text{Re}[w_N], \text{Im}[w_1], \dots, \text{Im}[w_N])^T, \end{aligned} \quad (10)$$

$$\tilde{\nabla}l(\mathbf{z}, y, \mathbf{v}) \stackrel{\text{def}}{=} \mathbf{G}^{-1}(\mathbf{v}) \cdot \nabla l(\mathbf{z}, y, \mathbf{v}). \quad (11)$$

Eq. (11) is the natural gradient of $l(\mathbf{z}, y, \mathbf{v})$, and the usual gradient $\nabla l(\mathbf{z}, y, \mathbf{v})$ is given by

$$\begin{aligned} \nabla l(\mathbf{z}, y, \mathbf{v}) &\stackrel{\text{def}}{=} \left(\frac{\partial l(\mathbf{z}, y, \mathbf{v})}{\partial v_1}, \dots, \frac{\partial l(\mathbf{z}, y, \mathbf{v})}{\partial v_{2N}} \right)^T \\ &= \left(\frac{\partial l(\mathbf{z}, y, \mathbf{w})}{\partial \text{Re}[w_1]}, \dots, \frac{\partial l(\mathbf{z}, y, \mathbf{w})}{\partial \text{Re}[w_N]}, \right. \\ &\quad \left. \frac{\partial l(\mathbf{z}, y, \mathbf{w})}{\partial \text{Im}[w_1]}, \dots, \frac{\partial l(\mathbf{z}, y, \mathbf{w})}{\partial \text{Im}[w_N]} \right)^T. \end{aligned} \quad (12)$$

The Riemannian metric tensor $\mathbf{G}(\mathbf{v})$ is the Fisher information matrix [3], and is given by

$$\mathbf{G}(\mathbf{v}) = (g_{ij}(\mathbf{v})), \quad (13)$$

$$g_{ij}(\mathbf{v}) = E \left[\frac{\partial \log p(\mathbf{z}, y; \mathbf{v})}{\partial v_i} \cdot \frac{\partial \log p(\mathbf{z}, y; \mathbf{v})}{\partial v_j} \right]. \quad (14)$$

B. Natural Gradient Learning in a Single Complex-Valued Neuron

In this section, the natural gradient descent learning algorithm for a single complex-valued neuron is given.

Consider a stochastic complex-valued neuron with N -inputs, weights $w_k = u_k + iv_k \in \mathbf{C}$ ($1 \leq k \leq N$), and a threshold value $\gamma = c + id \in \mathbf{C}$. Then, for N input signals $z_k = x_k + iy_k \in \mathbf{C}$ ($1 \leq k \leq N$), the stochastic complex-valued neuron generates

$$\begin{aligned} y &= f_C \left(\sum_{k=1}^N w_k z_k + \gamma \right) + n \\ &= X + iY \end{aligned} \quad (15)$$

where $f_C : \mathbf{C} \rightarrow \mathbf{C}$ is a so-called *split-type* complex-valued activation function which is defined to be

$$f_C(a + ib) = \varphi(a) + i\varphi(b) \quad (16)$$

for any $a + ib \in \mathbf{C}$, and $\varphi : \mathbf{R} \rightarrow \mathbf{R}$ is suitably chosen, for example, the sigmoid function

$$\varphi(s) = \frac{1}{1 + e^{-s}} \quad (17)$$

was used in [11], and the scaled error function

$$\varphi(s) = \frac{2}{\sqrt{\pi}} \int_0^{\frac{s}{\sqrt{2}}} e^{-t^2} dt \quad (18)$$

was used in [13]. $n = n_R + in_I$ is a complex-valued random variable subject to the complex normal distribution (or bivariate normal distribution) $N(\mu, \Sigma)$ where

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad (19)$$

$$\Sigma = \begin{bmatrix} \sigma & 0 \\ 0 & \sigma \end{bmatrix}. \quad (20)$$

We assume that the input signal $\mathbf{z} = \mathbf{x} + i\mathbf{y}$ where $\mathbf{x} \stackrel{\text{def}}{=} (x_1, \dots, x_N)^T, \mathbf{y} \stackrel{\text{def}}{=} (y_1, \dots, y_N)^T$ is subject to the multivariate complex normal distribution (or $2N$ -dimensional normal distribution) $N(\mathbf{0}, \mathbf{I})$ where the variance covariance matrix \mathbf{I} is the unit matrix; denote its joint probability density function by $q(\mathbf{z})$. The loss function $l(\mathbf{z}, y; \theta)$ is defined as

$$l(\mathbf{z}, y; \theta) \stackrel{\text{def}}{=} -\log p(\mathbf{z}, y; \theta) \\ = -\log q(\mathbf{z}) - \log p(y|\mathbf{z}; \theta), \quad (21)$$

where $\theta \stackrel{\text{def}}{=} (\mathbf{u}^T, \mathbf{v}^T, c, d)^T, \mathbf{u} = (u_1, \dots, u_N)^T, \mathbf{v} = (v_1, \dots, v_N)^T$.

Given the sequence of the complex-valued training signals $\{(\mathbf{z}_t, y_t) \in \mathbf{C}^N \times \mathbf{C}, t = 1, 2, \dots\}$, the the natural gradient descent algorithm for the stochastic complex-valued neuron is given by

$$\theta_{t+1} = \theta_t - \varepsilon_t \cdot \mathbf{G}^{-1}(\theta_t) \cdot \nabla l(\mathbf{z}_t, y_t; \theta_t). \quad (22)$$

We shall calculate the Fisher information matrix $\mathbf{G}(\theta) = (g_{ij}(\theta))$. For any $1 \leq i, j \leq 2N + 2$,

$$g_{ij}(\theta) = E \left[\frac{\partial \log p(\mathbf{z}, y; \theta)}{\partial \theta_i} \cdot \frac{\partial \log p(\mathbf{z}, y; \theta)}{\partial \theta_j} \right] \\ = E \left[\frac{\partial \log p(y|\mathbf{z}; \theta)}{\partial \theta_i} \cdot \frac{\partial \log p(y|\mathbf{z}; \theta)}{\partial \theta_j} \right]. \quad (23)$$

(from Eq. (21))

Here, since n_R is independent of n_I ,

$$\log p(y|\mathbf{z}; \theta) = \log p((\varphi(S) + n_R) + i(\varphi(T) + n_I)|\mathbf{z}; \theta) \\ = \log p(\varphi(S) + n_R|\mathbf{z}; \theta) \\ + \log p(\varphi(T) + n_I|\mathbf{z}; \theta) \\ = \log p(X|\mathbf{z}; \theta) + \log p(Y|\mathbf{z}; \theta), \quad (24)$$

where $S = \text{Re} \left[\sum_{k=1}^N w_k z_k + \gamma \right], T = \text{Im} \left[\sum_{k=1}^N w_k z_k + \gamma \right]$. Thus, for any $1 \leq i, j \leq 2N + 2$,

$$g_{ij}(\theta) = E \left[\frac{\partial \log p(X|\mathbf{z}; \theta)}{\partial \theta_i} \cdot \frac{\partial \log p(X|\mathbf{z}; \theta)}{\partial \theta_j} \right] \\ + 2E \left[\frac{\partial \log p(X|\mathbf{z}; \theta)}{\partial \theta_i} \cdot \frac{\partial \log p(Y|\mathbf{z}; \theta)}{\partial \theta_j} \right] \\ + E \left[\frac{\partial \log p(Y|\mathbf{z}; \theta)}{\partial \theta_i} \cdot \frac{\partial \log p(Y|\mathbf{z}; \theta)}{\partial \theta_j} \right]. \quad (25)$$

By simple calculations, we obtain

$$g_{ij}(\theta) = \frac{1}{\sigma^2} \{ E[(\varphi'(S))^2 x_i x_j] \\ + E[(\varphi'(T))^2 y_i y_j] \} \\ (1 \leq i, j \leq N), \quad (26)$$

$$g_{ij}(\theta) = \frac{1}{\sigma^2} \{ E[(\varphi'(S))^2 (-y_{i-N}) x_j] \\ + E[(\varphi'(T))^2 x_{i-N} y_j] \} \\ = g_{ji}(\theta) \\ (N + 1 \leq i \leq 2N, 1 \leq j \leq N), \quad (27)$$

$$g_{ij}(\theta) = \frac{1}{\sigma^2} \{ E[(\varphi'(S))^2 y_{i-N} y_{j-N}] \\ + E[(\varphi'(T))^2 x_{i-N} x_{j-N}] \} \\ (N + 1 \leq i, j \leq 2N), \quad (28)$$

$$g_{2N+1,j}(\theta) = \frac{1}{\sigma^2} E[(\varphi'(S))^2 x_j] \\ = g_{j,2N+1}(\theta) \quad (1 \leq j \leq N), \quad (29)$$

$$g_{2N+1,j}(\theta) = \frac{1}{\sigma^2} E[(\varphi'(S))^2 (-y_{j-N})] \\ (N + 1 \leq j \leq 2N), \quad (30)$$

$$g_{2N+2,j}(\theta) = \frac{1}{\sigma^2} E[(\varphi'(T))^2 y_j] \quad (1 \leq j \leq N), \quad (31)$$

$$g_{2N+2,j}(\theta) = \frac{1}{\sigma^2} E[(\varphi'(T))^2 x_{j-N}] \\ (N + 1 \leq j \leq 2N), \quad (32)$$

$$g_{2N+1,2N+1}(\theta) = \frac{1}{\sigma^2} E[(\varphi'(S))^2], \quad (33)$$

$$g_{2N+2,2N+1}(\theta) = 0 \\ = g_{2N+1,2N+2}(\theta), \quad (34)$$

$$g_{2N+2,2N+2}(\theta) = \frac{1}{\sigma^2} E[(\varphi'(T))^2]. \quad (35)$$

From Eqs. (26) – (35), we can rewrite $\mathbf{G}(\theta)$ as

$$\mathbf{G}(\theta) = \frac{1}{\sigma^2} \mathbf{A}(\theta), \quad (36)$$

where

$$\mathbf{A}(\theta) \stackrel{\text{def}}{=} \left(\begin{array}{cc|cc} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{a}_{13} & \mathbf{a}_{14} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{a}_{23} & \mathbf{a}_{24} \\ \hline \mathbf{a}_{31} & \mathbf{a}_{32} & & \\ \mathbf{a}_{41} & \mathbf{a}_{42} & & \mathbf{A}_{44} \end{array} \right), \quad (37)$$

$\mathbf{A}_{11} = \sigma^2 \cdot \text{Eq. (26)}, \mathbf{A}_{21} = \mathbf{A}_{12}^T = \sigma^2 \cdot \text{Eq. (27)}, \mathbf{A}_{22} = \sigma^2 \cdot \text{Eq. (28)}, \mathbf{a}_{13} = \mathbf{a}_{31}^T = \sigma^2 \cdot \text{Eq. (29)}, \mathbf{a}_{32} = \mathbf{a}_{23}^T = \sigma^2 \cdot \text{Eq. (30)}, \mathbf{a}_{14} = \mathbf{a}_{41}^T = \sigma^2 \cdot \text{Eq. (31)}, \mathbf{a}_{42} = \mathbf{a}_{24}^T = \sigma^2 \cdot \text{Eq. (32)},$ and

$$\mathbf{A}_{44} = \left(\begin{array}{cc} \sigma^2 \cdot \text{Eq. (33)} & \text{Eq. (34)} \\ \text{Eq. (34)} & \sigma^2 \cdot \text{Eq. (35)} \end{array} \right). \quad (38)$$

In what follows, each submatrix of $\mathbf{A}(\theta)$ (Eq. (37)) is calculated. Let $u = \|\mathbf{u}\| = \sqrt{u_1^2 + \dots + u_N^2}, v = \|\mathbf{v}\| = \sqrt{v_1^2 + \dots + v_N^2}, \mathbf{u}_1 = \mathbf{u}/u, \mathbf{v}_1 = \mathbf{v}/v$, and extend $\mathbf{u}_1, \mathbf{v}_1$ to orthonormal bases $\{\mathbf{u}_1, \dots, \mathbf{u}_N\}, \{\mathbf{v}_1, \dots, \mathbf{v}_N\}$ for \mathbf{R}^N , respectively. Then, the real part $\mathbf{x} \in \mathbf{R}^N$ and the imaginary

part $\mathbf{y} \in \mathbf{R}^N$ of the input signal $\mathbf{z} = \mathbf{x} + i\mathbf{y} \in \mathbf{C}^N$ can be decomposed as

$$\mathbf{x} = \sum_{i=1}^N \chi_i \mathbf{u}_i, \quad (39)$$

$$\mathbf{y} = \sum_{i=1}^N \psi_i \mathbf{v}_i. \quad (40)$$

Then, noticing that $\chi_i, \psi_i \sim N(0, 1)$, we have

$$\begin{aligned} E[\chi_i \chi_j] &= E[\mathbf{u}_i^T \mathbf{x} \mathbf{x}^T \mathbf{u}_j] = \mathbf{u}_i^T E[\mathbf{x} \mathbf{x}^T] \mathbf{u}_j \\ &= \mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}, \end{aligned} \quad (41)$$

$$\begin{aligned} \mathbf{x} \mathbf{x}^T &= \chi_1^2 \mathbf{u}_1 \mathbf{u}_1^T + \sum_{i=2}^N \chi_1 \chi_i (\mathbf{u}_1 \mathbf{u}_i^T + \mathbf{u}_i \mathbf{u}_1^T) \\ &\quad + \sum_{i,j=2}^N \chi_i \chi_j \mathbf{u}_i \mathbf{u}_j^T, \end{aligned} \quad (42)$$

$$\mathbf{u}^T \cdot \mathbf{x} = \mathbf{u}^T (\chi_1 \mathbf{u}_1 + \dots + \chi_N \mathbf{u}_N) = \chi_1 u, \quad (43)$$

$$\mathbf{v}^T \cdot \mathbf{y} = \mathbf{v}^T (\psi_1 \mathbf{v}_1 + \dots + \psi_N \mathbf{v}_N) = \psi_1 v. \quad (44)$$

From Eqs. (41) – (44), we find that the first term of Eq. (26) (\mathbf{A}_{11}) is given by

$$\begin{aligned} E[(\varphi'(S))^2 \mathbf{x} \mathbf{x}^T] &= E[(\varphi'(\mathbf{u}^T \mathbf{x} - \mathbf{v}^T \mathbf{y} + c))^2 \mathbf{x} \mathbf{x}^T] \\ &= E[(\varphi'(\chi_1 u - \psi_1 v + c))^2 \chi_1^2] \mathbf{u}_1 \mathbf{u}_1^T \\ &\quad + E[(\varphi'(\chi_1 u - \psi_1 v + c))^2] \sum_{i=2}^N \mathbf{u}_i \mathbf{u}_i^T. \end{aligned} \quad (45)$$

Next, in order to calculate the second term of Eq. (26) (\mathbf{A}_{11}), decompose the real part $\mathbf{x} \in \mathbf{R}^N$ and the imaginary part $\mathbf{y} \in \mathbf{R}^N$ of the input signal $\mathbf{z} = \mathbf{x} + i\mathbf{y} \in \mathbf{C}^N$ as

$$\mathbf{x} = \sum_{i=1}^N \chi'_i \mathbf{v}_i, \quad (46)$$

$$\mathbf{y} = \sum_{i=1}^N \psi'_i \mathbf{u}_i. \quad (47)$$

Then, we have

$$E[\psi'_i \psi'_j] = \delta_{ij}, \quad (48)$$

$$\begin{aligned} \mathbf{y} \mathbf{y}^T &= (\psi'_1)^2 \mathbf{u}_1 \mathbf{u}_1^T + \sum_{i=2}^N \psi'_1 \psi'_i (\mathbf{u}_1 \mathbf{u}_i^T + \mathbf{u}_i \mathbf{u}_1^T) \\ &\quad + \sum_{i,j=2}^N \psi'_i \psi'_j \mathbf{u}_i \mathbf{u}_j^T, \end{aligned} \quad (49)$$

$$\mathbf{v}^T \cdot \mathbf{x} = \mathbf{v}^T (\chi'_1 \mathbf{v}_1 + \dots + \chi'_N \mathbf{v}_N) = \chi'_1 v, \quad (50)$$

$$\mathbf{u}^T \cdot \mathbf{y} = \mathbf{u}^T (\psi'_1 \mathbf{u}_1 + \dots + \psi'_N \mathbf{u}_N) = \psi'_1 u. \quad (51)$$

From Eqs. (48) – (51), we find that the second term of Eq.

(26) (\mathbf{A}_{11}) is given by

$$\begin{aligned} E[(\varphi'(T))^2 \mathbf{y} \mathbf{y}^T] &= E[(\varphi'(\mathbf{v}^T \mathbf{x} + \mathbf{u}^T \mathbf{y} + d))^2 \mathbf{y} \mathbf{y}^T] \\ &= E[(\varphi'(\chi'_1 v + \psi'_1 u + d))^2 (\psi'_1)^2] \mathbf{u}_1 \mathbf{u}_1^T \\ &\quad + E[(\varphi'(\chi'_1 v + \psi'_1 u + d))^2] \sum_{i=2}^N \mathbf{u}_i \mathbf{u}_i^T. \end{aligned} \quad (52)$$

Thus, from Eqs. (45) and (52),

$$\begin{aligned} \mathbf{A}_{11} &= E[(\varphi'(S))^2 \mathbf{x} \mathbf{x}^T] + E[(\varphi'(T))^2 \mathbf{y} \mathbf{y}^T] \\ &= \{E[(\varphi'(\chi_1 u - \psi_1 v + c))^2 \chi_1^2] \\ &\quad + E[(\varphi'(\chi'_1 v + \psi'_1 u + d))^2 \psi_1^2]\} \mathbf{u}_1 \mathbf{u}_1^T \\ &\quad + \{E[(\varphi'(\chi_1 u - \psi_1 v + c))^2] \\ &\quad + E[(\varphi'(\chi'_1 v + \psi'_1 u + d))^2]\} \sum_{i=2}^N \mathbf{u}_i \mathbf{u}_i^T \\ &= d_0(u, v, c, d) \cdot I \\ &\quad + \{d_2(u, v, c, d) - d_0(u, v, c, d)\} \cdot \frac{\mathbf{u} \mathbf{u}^T}{u^2}, \end{aligned} \quad (53)$$

where

$$d_0(u, v, c, d) \stackrel{\text{def}}{=} E[\{\varphi'(\chi_1 u - \psi_1 v + c)\}^2] + E[\{\varphi'(\chi'_1 v + \psi'_1 u + d)\}^2], \quad (54)$$

$$d_2(u, v, c, d) \stackrel{\text{def}}{=} E[\{\varphi'(\chi_1 u - \psi_1 v + c)\}^2 \chi_1^2] + E[\{\varphi'(\chi'_1 v + \psi'_1 u + d)\}^2 (\psi'_1)^2]. \quad (55)$$

Similarly, we have

$$\begin{aligned} \mathbf{A}_{21} &= d_{11}(u, v, c, d) \cdot \frac{\mathbf{v} \mathbf{u}^T}{uv} \\ &= \mathbf{A}_{12}^T, \end{aligned} \quad (56)$$

$$\begin{aligned} \mathbf{A}_{22} &= d_0(u, v, c, d) \cdot I \\ &\quad + \{d'_2(u, v, c, d) - d_0(u, v, c, d)\} \cdot \frac{\mathbf{v} \mathbf{v}^T}{v^2}, \end{aligned} \quad (57)$$

$$\begin{aligned} \mathbf{a}_{31} &= d_{1x}(u, v, c) \cdot \mathbf{u}_1^T \\ &= \mathbf{a}_{13}^T, \end{aligned} \quad (58)$$

$$\begin{aligned} \mathbf{a}_{32} &= -d_{1y}(u, v, c) \cdot \mathbf{v}_1^T \\ &= \mathbf{a}_{23}^T, \end{aligned} \quad (59)$$

$$\begin{aligned} \mathbf{a}_{41} &= d'_{1y}(u, v, d) \cdot \mathbf{u}_1^T \\ &= \mathbf{a}_{14}^T, \end{aligned} \quad (60)$$

$$\begin{aligned} \mathbf{a}_{42} &= d'_{1x}(u, v, d) \cdot \mathbf{v}_1^T \\ &= \mathbf{a}_{24}^T, \end{aligned} \quad (61)$$

$$\mathbf{A}_{44} = \begin{pmatrix} d_0(u, v, c) & 0 \\ 0 & d'_0(u, v, d) \end{pmatrix} \quad (62)$$

where

$$d_{11}(u, v, c, d) \stackrel{\text{def}}{=} E[\{\varphi'(\chi_1'v + \psi_1'u + d)\}^2 \chi_1' \psi_1'] - E[\{\varphi'(\chi_1u - \psi_1v + c)\}^2 \chi_1 \psi_1], \quad (63)$$

$$d_2'(u, v, c, d) \stackrel{\text{def}}{=} E[\{\varphi'(\chi_1u - \psi_1v + c)\}^2 \psi_1^2] + E[\{\varphi'(\chi_1'v + \psi_1'u + d)\}^2 (\chi_1')^2], \quad (64)$$

$$d_{1x}(u, v, c) \stackrel{\text{def}}{=} E[\{\varphi'(\chi_1u - \psi_1v + c)\}^2 \chi_1], \quad (65)$$

$$d_{1y}(u, v, c) \stackrel{\text{def}}{=} E[\{\varphi'(\chi_1u - \psi_1v + c)\}^2 \psi_1], \quad (66)$$

$$d_{1y}'(u, v, d) \stackrel{\text{def}}{=} E[\{\varphi'(\chi_1'v + \psi_1'u + d)\}^2 \psi_1], \quad (67)$$

$$d_{1x}'(u, v, d) \stackrel{\text{def}}{=} E[\{\varphi'(\chi_1'v + \psi_1'u + d)\}^2 \chi_1], \quad (68)$$

$$d_0(u, v, c) \stackrel{\text{def}}{=} E[\{\varphi'(\chi_1u - \psi_1v + c)\}^2], \quad (69)$$

$$d_0'(u, v, d) \stackrel{\text{def}}{=} E[\{\varphi'(\chi_1'v + \psi_1'u + d)\}^2]. \quad (70)$$

We compute the inverse of $A(\theta)$ (Eq. (37)) using the following formula used in [13].

Lemma 1:

$$\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}^{-1} = \begin{pmatrix} B^{11} & B^{12} \\ B^{21} & B^{22} \end{pmatrix}, \quad (71)$$

provided

$$|B_{11}| \neq 0, |B_{22} - B_{21}B_{11}^{-1}B_{12}| \neq 0 \quad (72)$$

where

$$B^{11} = B_{11}^{-1} + B_{11}^{-1}B_{12}B_{22,1}^{-1}B_{21}B_{11}^{-1}, \quad (73)$$

$$B_{22,1} = B_{22} - B_{21}B_{11}^{-1}B_{12}, \quad (74)$$

$$B^{22} = B_{22,1}^{-1}, \quad (75)$$

$$B^{12} = (B^{21})^T = -B_{11}^{-1}B_{12}B_{22,1}^{-1}. \quad (76)$$

□

By using Lemma 1, we have

$$A(\theta)^{-1} = \begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix}, \quad (77)$$

where

$$A^{11} = \begin{pmatrix} a^{11} & a^{12} \\ a^{21} & a^{22} \end{pmatrix}, \quad (78)$$

$$a^{11} = \frac{I}{d_0} + \left\{ G + \left(\frac{1}{d_0} + G \right) (m_1l_1 + m_2l_2) + H(m_1l_3 + m_2l_4) \right\} \frac{uu^T}{u^2}, \quad (79)$$

$$a^{12} = \left\{ H + H(m_1l_1 + m_2l_2) + \left(\frac{1}{d_0} + E \right) (m_1l_3 + m_2l_4) \right\} \frac{uv^T}{uv}, \quad (80)$$

$$a^{21} = \left\{ H + \left(\frac{1}{d_0} + G \right) (m_3l_1 + m_4l_2) + H(m_3l_3 + m_4l_4) \right\} \frac{vu^T}{vu}, \quad (81)$$

$$a^{22} = \frac{I}{d_0} + \{ E + H(m_3l_1 + m_4l_2) + \left(\frac{1}{d_0} + E \right) (m_3l_3 + m_4l_4) \} \frac{vv^T}{v^2}, \quad (82)$$

$$A^{12} = \begin{pmatrix} l_1u/u & l_2u/u \\ l_3v/v & l_4v/v \end{pmatrix} = (A^{21})^T, \quad (83)$$

$$A^{22} = \frac{1}{k_1k_2 - k^2} \begin{pmatrix} k_2 & -k \\ -k & k_1 \end{pmatrix}, \quad (84)$$

$$E = \frac{d_2}{d_2d_2' - d_{11}^2} - \frac{1}{d_0}, \quad (85)$$

$$F = \frac{1}{d_2} - \frac{1}{d_0}, \quad (86)$$

$$G = Ed_{11}^2 \left(\frac{1 + 2F}{d_0} + F^2 \right) + \frac{d_{11}^2}{d_0} \left(F^2 + 2F + \frac{1}{d_0^2} \right) + F, \quad (87)$$

$$H = -d_{11} \left(\frac{1}{d_0} + E \right) \left(\frac{1}{d_0} + F \right), \quad (88)$$

$$k = -d_{1x}d_{1y}' \left(\frac{1}{d_0} + G \right) - H(d_{1y}d_{1y}' - d_{1x}d_{1x}') + d_{1x}'d_{1y} \left(\frac{1}{d_0} + E \right), \quad (89)$$

$$k_1 = d_0 - d_{1x}^2 \left(\frac{1}{d_0} + G \right) + 2Hd_{1x}d_{1y}' - d_{1y}^2 \left(\frac{1}{d_0} + E \right), \quad (90)$$

$$k_2 = d_0' - d_{1y}'^2 \left(\frac{1}{d_0} + G \right) - 2Hd_{1x}'d_{1y}' - d_{1x}'^2 \left(\frac{1}{d_0} + E \right), \quad (91)$$

$$l_1 = \frac{1}{k_1k_2 - k^2} \left\{ -k_2 \left(\frac{d_{1x}}{d_0} + Gd_{1x} - Hd_{1y} \right) + k \left(\frac{d_{1y}'}{d_0} + Gd_{1y}' + Hd_{1x}' \right) \right\}, \quad (92)$$

ACKNOWLEDGMENT

The author would like to give special thanks to the anonymous reviewers for valuable comments.

REFERENCES

- [1] S. Amari, *Differential-geometrical methods in statistics*, Lecture Notes in Statistics vol. 28, Springer-Verlag, 1985.
- [2] S. Amari, "Neural learning in structured parameter spaces – Natural Riemannian gradient," In M. C. Mozer, M. I. Jordan, & Th. Petsche (Eds.), *Advances in neural processing systems*, 9, Cambridge, MA: MIT Press, 1996.
- [3] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.
- [4] S. Amari, H. Park, and T. Ozeki, "Singularities affect dynamics of learning in neuromanifolds," *Neural Computation*, vol. 18, no. 5, pp. 1007–1065, 2006.
- [5] G. M. Georgiou and C. Koutsougeras, "Complex domain backpropagation," *IEEE Trans. Circuits and Systems-II: Analog and Digital Signal Processing*, vol. 39, no. 5, pp. 330–334, 1992.
- [6] T. Nitta (Ed.), *Complex-valued neural networks: utilizing high-dimensional parameters*, Information Science Reference, Pennsylvania, USA, 2009.
- [7] T. Nitta, "Local minima in hierarchical structures of complex-valued neural networks," *Neural Networks*, vol. 43, pp. 1–7, 2013.
- [8] T. Nitta, "Plateau in a polar variable complex-valued neuron," Proceedings of the 6th International Conference on Agents and Artificial Intelligence, ICAART2014-Anger, March 6-8, 2014, pp. 526–531.
- [9] A. Hirose (Ed.), *Complex-valued neural networks: advances and applications in the IEEE press series on computational intelligence*, Wiley-IEEE Press, 2013.
- [10] M. S. Kim and C. C. Guest, "Modification of backpropagation networks for complex-valued signal processing in frequency domain," *Proc. International Joint Conference on Neural Networks*, San Diego, June, 1990, vol. 3, pp. 27–31.
- [11] T. Nitta, "An extension of the back-propagation algorithm to complex numbers," *Neural Networks*, vol. 10, no. 8, pp. 1392–1415, 1997.
- [12] T. Nitta, "Orthogonality of decision boundaries in complex-valued neural networks," *Neural Computation*, vol. 16, no. 1, pp. 73–97, 2004.
- [13] H. H. Yang and S. Amari, "Complexity issues in natural gradient descent method for training multilayer perceptrons," *Neural Computation*, vol. 10, no. 8, pp. 2137–2157, 1998.

V. CONCLUDING REMARK

We have developed the natural gradient descent method for the multilayer stochastic complex-valued neural networks, and derived the natural gradient for a single stochastic complex-valued neuron. Since we assume that the variance σ_1^2 for the real part of the complex-valued additive noise is equal to that σ_2^2 for the imaginary, the Riemannian metric tensor is given by $\mathbf{G}(\theta) = (1/\sigma^2)\mathbf{A}(\theta)$ where $\mathbf{A}(\theta)$ does not depend on $\sigma^2 = \sigma_1^2 = \sigma_2^2$. The situation is, however, not the same if $\sigma_1 \neq \sigma_2$. And also, the Riemannian metric tensor will be more complicated if the covariance is not zero.

In future studies, based on the results of this paper, we will derive the natural gradient descent algorithm for the three-layered stochastic complex-valued neural network, and make clear its properties via computer simulations. Since there exist many singular points in the three-layered stochastic complex-valued neural network [7], it is expected that the complex-valued natural gradient method improves its learning performance dramatically. And also, it has been shown that there exist singular points in the polar variable complex-valued neurons [8]. We will apply the natural gradient method to the polar variable complex-valued neurons.