

IJCSIS Vol. 15 No. 11, November 2017
ISSN 1947-5500

International Journal of Computer Science & Information Security

© IJCSIS PUBLICATION 2017
Pennsylvania, USA

Indexed and technically co-sponsored by :



Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS

International Journal of Computer Science and Information Security (IJCSIS) January-December 2017 Issues

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.

See authors guide for manuscript preparation and submission guidelines.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Scopus Database, Cornell University Library, ScientificCommons, ProQuest, EBSCO and more.

Deadline: see web site

Notification: see web site

Revision: see web site

Publication: see web site

Context-aware systems

Networking technologies

Security in network, systems, and applications

Evolutionary computation

Industrial systems

Evolutionary computation

Autonomic and autonomous systems

Bio-technologies

Knowledge data systems

Mobile and distance education

Intelligent techniques, logics and systems

Knowledge processing

Information technologies

Internet and web technologies

Digital information processing

Cognitive science and knowledge

Agent-based systems

Mobility and multimedia systems

Systems performance

Networking and telecommunications

Software development and deployment

Knowledge virtualization

Systems and networks on the chip

Knowledge for global defense

Information Systems [IS]

IPv6 Today - Technology and deployment

Modeling

Software Engineering

Optimization

Complexity

Natural Language Processing

Speech Synthesis

Data Mining

For more topics, please see web site <https://sites.google.com/site/ijcsis/>

 arXiv.org  Google scholar

 SCIRUS
search engine for science

 ScientificCommons

 Scribd

 .docstoc
find and share professional documents

 BASE
Bielefeld Academic Search Engine

 CiteSeerX beta

 Computer Science
Bibliography

 DOAJ
DIRECTORY OF
OPEN ACCESS
JOURNALS

 EBSCO
HOST

 ProQuest

For more information, please visit the journal website (<https://sites.google.com/site/ijcsis/>)

Editorial Message from Editorial Board

It is our great pleasure to present the November 2017 issue (Volume 15 Number 11, Part I & II) of the International Journal of Computer Science and Information Security (IJCSIS). High quality research, survey & review articles are proposed from experts in the field, promoting insight and understanding of the state of the art, and trends in computer science and technology. It especially provides a platform for high-caliber academics, practitioners and PhD/Doctoral graduates to publish completed work and latest research outcomes. According to Google Scholar, up to now papers published in IJCSIS have been cited over 9775 times and the number is quickly increasing. This statistics shows that IJCSIS has established the first step to be an international and prestigious journal in the field of Computer Science and Information Security. There have been many improvements to the processing of papers; we have also witnessed a significant growth in interest through a higher number of submissions as well as through the breadth and quality of those submissions. IJCSIS is indexed in major academic/scientific databases and important repositories, such as: Google Scholar, Thomson Reuters, ArXiv, CiteSeerX, Cornell's University Library, Ei Compendex, ISI Scopus, DBLP, DOAJ, ProQuest, ResearchGate, Academia.edu and EBSCO among others.

On behalf of IJCSIS community and the sponsors, we congratulate the authors and thank the reviewers for their outstanding efforts to review and recommend high quality papers for publication. In particular, we would like to thank the international academia and researchers for continued support by citing papers published in IJCSIS. Without their sustained and unselfish commitments, IJCSIS would not have achieved its current premier status.

"We support researchers to succeed by providing high visibility & impact value, prestige and excellence in research publication." For further questions or other suggestions please do not hesitate to contact us at ijcsiseditor@gmail.com.

*A complete list of journals can be found at:
<http://sites.google.com/site/ijcsis/>*

IJCSIS Vol. 15, No. 11, November 2017 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):



Open Access This Journal is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source.



Bibliographic Information

ISSN: 1947-5500

Monthly publication (Regular Special Issues)

Commenced Publication since May 2009

Editorial / Paper Submissions:

IJCSIS Managing Editor

[\(ijcsiseditor@gmail.com\)](mailto:ijcsiseditor@gmail.com)

Pennsylvania, USA

Tel: +1 412 390 5159

IJCSIS EDITORIAL BOARD

IJCSIS Editorial Board	IJCSIS Guest Editors / Associate Editors
Dr. Shimon K. Modi [Profile] Director of Research BSPA Labs, Purdue University, USA	Dr Riktesh Srivastava [Profile] Associate Professor, Information Systems, Skyline University College, Sharjah, PO 1797, UAE
Professor Ying Yang, PhD. [Profile] Computer Science Department, Yale University, USA	Dr. Jianguo Ding [Profile] Norwegian University of Science and Technology (NTNU), Norway
Professor Hamid Reza Naji, PhD. [Profile] Department of Computer Engineering, Shahid Beheshti University, Tehran, Iran	Dr. Naseer Alquraishi [Profile] University of Wasit, Iraq
Professor Yong Li, PhD. [Profile] School of Electronic and Information Engineering, Beijing Jiaotong University, P. R. China	Dr. Kai Cong [Profile] Intel Corporation, & Computer Science Department, Portland State University, USA
Professor Mokhtar Beldjehem, PhD. [Profile] Sainte-Anne University, Halifax, NS, Canada	Dr. Omar A. Alzubi [Profile] Al-Balqa Applied University (BAU), Jordan
Professor Yousef Farhaoui, PhD. Department of Computer Science, Moulay Ismail University, Morocco	Dr. Jorge A. Ruiz-Vanoye [Profile] Universidad Autónoma del Estado de Morelos, Mexico
Dr. Alex Pappachen James [Profile] Queensland Micro-nanotechnology center, Griffith University, Australia	Prof. Ning Xu, Wuhan University of Technology, China
Professor Sanjay Jasola [Profile] Gautam Buddha University	Dr . Bilal Alatas [Profile] Department of Software Engineering, Firat University, Turkey
Dr. Siddhivinayak Kulkarni [Profile] University of Ballarat, Ballarat, Victoria, Australia	Dr. Ioannis V. Koskosas, University of Western Macedonia, Greece
Dr. Reza Ebrahimi Atani [Profile] University of Guilan, Iran	Dr Venu Kuthadi [Profile] University of Johannesburg, Johannesburg, RSA
Dr. Dong Zhang [Profile] University of Central Florida, USA	Dr. Zhihan Lv [Profile] Chinese Academy of Science, China
Dr. Vahid Esmaeilzadeh [Profile] Iran University of Science and Technology	Prof. Ghulam Qasim [Profile] University of Engineering and Technology, Peshawar, Pakistan
Dr. Jiliang Zhang [Profile] Northeastern University, China	Prof. Dr. Maqbool Uddin Shaikh [Profile] Preston University, Islamabad, Pakistan
Dr. Jacek M. Czerniak [Profile] Casimir the Great University in Bydgoszcz, Poland	Dr. Musa Peker [Profile] Faculty of Technology, Mugla Sitki Kocman University, Turkey
Dr. Binh P. Nguyen [Profile] National University of Singapore	Dr. Wencan Luo [Profile] University of Pittsburgh, US
Professor Seifeidne Kadry [Profile] American University of the Middle East, Kuwait	Dr. Ijaz Ali Shoukat [Profile] King Saud University, Saudi Arabia
Dr. Riccardo Colella [Profile] University of Salento, Italy	Dr. Yilun Shang [Profile] Tongji University, Shanghai, China
Dr. Sedat Akylek [Profile] Ondokuz Mayis University, Turkey	Dr. Sachin Kumar [Profile] Indian Institute of Technology (IIT) Roorkee

Dr Basit Shahzad [Profile] King Saud University, Riyadh - Saudi Arabia	
Dr. Sherzod Turaev [Profile] International Islamic University Malaysia	

ISSN 1947 5500 Copyright © IJCSIS, USA.

TABLE OF CONTENTS

1. PaperID 31101704: Extended Multi Queue Job Scheduling in Cloud (pp. 1-8)

Neeraj Kumar Pandey, Research Scholar, Uttarakhand University, Dehradun Uttarakhand

Sumit Chaudhary, Research Scholar, Uttarakhand University, Dehradun Uttarakhand

N. K. Joshi, Director, Uttarakhand Institute of Technology, Uttarakhand University, Dehradun Uttarakhand

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

2. PaperID 31101705: Analyzing and Predicting Hash tags Popularity on Twitter (pp. 9-16)

Katarla Krishna Vasu (1) & Dr. Vedula Venkateswara Rao (2)

(1) Dept of CSE, Sri Vasavi Engineering College Pedatadepalli, Tadepalligudem, A.P, India

(2) Associate Professor, Dept of CSE, Sri Vasavi Engineering College Pedatadepalli, Tadepalligudem, A.P, India

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

3. PaperID 31101706: Invisible Digital Watermarking Methodology for Image Validation (pp. 17-22)

Dr. V. Kakulapati, SNIST, Hyderabad, India

Dr. SR Kattamuri, SNIST, Hyderabad, India

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

4. PaperID 31101707: Designing an Efficient Distributed Algorithm for Big Data Analytics: Issues and Challenges (pp. 23-28)

Mohammed S. Al-kahtani, Dept. of Computer Engg., Prince Sattam bin Abdulaziz University, Saudi Arabia

Lutful Karim, School of ICT, Seneca College of Applied Arts & Technology, Toronto, Canada

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

5. PaperID 31101711: A Survey on Memory Virtualization in Cloud (pp. 29-34)

P. V. S. S. Gangadhar, Scientist-D, Research Scholar, NIC, MEITY, Govt of India, Gitam University, Visakhapatnam, Andhra Pradesh, India

Dr. Ashok Kumar Hota, Scientist-F, NIC, MEITY, Govt of India

Dr. Mandapati Venkateswara Rao, Professor, Dept of IT, Gitam Institute of Technology, Gitam University, Visakhapatnam, Andhra Pradesh, India

Dr. Vedula Venkateswara Rao, Professor, Dept of CSE, Sri Vasavi Engineering College, Tadepalligudem, Andhra Pradesh, India

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

6. PaperID 31101715: A Comparative Study on MAC Protocols for Wireless Sensor Networks on Energy Reduction (pp. 35-40)

Rajan Sharma, Research Scholar, Department of Electronics and Communication Engineering, I.K. Gujral Punjab Technical University, Jalandhar, Punjab, India
Balwinder Singh Sohi, Department of Electronics and Communication, CGC Group of Colleges, Punjab, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

7. PaperID 31101717: Performance Analysis of Adaptive Routing Protocol for Cognitive Radio Wireless Sensor Networks using Bio-inspired Methods (pp. 41-49)

(1) Amit N. Thakare, (2) Dr. Latesh Malik (Bhagat), (3) Mrs. Achamma Thomas
(1, 3) Department of Computer Science & Engineering, G.H. Raisoni College of Engineering, Nagpur, MS, India
(2) Department of Computer Science & Engineering, Government College of Engineering, Nagpur, MS, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

8. PaperID 31101722: Application of the ORBAC Model in the Context of Physical Access Control (pp. 50-63)

Belbergui Chaimaa, STIC Laboratory, Chouaib Doukkali University, El jadida, Morocco
Elkamoun Najib, STIC Laboratory, Chouaib Doukkali University, El jadida, Morocco
Rachid Hilal, STIC Laboratory, Chouaib Doukkali University, El jadida, Morocco

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

9. PaperID 31101723: Clustering Algorithms: A Review (pp. 64-76)

*Nikitha Johnsirani Venkatesan & Dong Ryeol Shin**
School of Electrical and Computer Engineering, Sungkyunkwan University Suwon, South Korea

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

10. PaperID 31101725: Implementation Challenges in Campus Network Security (pp. 77-81)

Zaka Ullah, Department of Computer Science, Lahore Garrison University, Lahore, Pakistan
Muhammad Zulkifl Hasan, Department of Computer Science, Lahore Garrison University, Lahore, Pakistan

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

11. PaperID 31101728: Improving the Data Quality in the Research Information Systems (pp. 82-86)

Otmane Azeroual (abc),
(a) German Center for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a, Berlin, 10117, Germany
(b) Otto-von-Guericke-University Magdeburg, Department of Computer Science, Institute for Technical and Business Information Systems Database Research Group, P.O. Box 4120; 39106 Magdeburg, Germany
(c) University of Applied Sciences HTW Berlin, Study Program Computational Science and Engineering, Wilhelminenhofstraße 75A, 12459 Berlin, Germany

Mohammad Abuosba (c),
(c) University of Applied Sciences HTW Berlin, Study Program Computational Science and Engineering, Wilhelminenhofstraße 75A, 12459 Berlin, Germany

[Full Text: PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

12. PaperID 31101733: One Hop Forwarding Technique for QOS Routing with an Improved Fault Tolerant Clustering Mechanism for Multimedia Sensor Networks (pp. 87-100)

(1) *R. Guru*, (2) *Dr. Siddaraju*, (3) *Ananda babu*

(1) *Research scholar*, (2) *Professor*, (3) *Assistant Professor*

(1) *Sri Jayachamarajendra College of Engineering, Mysuru, India.*

(2) *Dr. Ambedkar Institute of Technology, Bangalore, India*

(3) *Kalpataru Institute of Engineering, Tiptur, India.*

[Full Text: PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

13. PaperID 31101736: Constructions of Design Elements from Software Requirement Specifications (SRS) (pp. 101-105)

Syed Naimatullah Hussain, CIT, Taif University, Taif, KSA

[Full Text: PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

14. PaperID 31101737: An Efficient Way of Medical Image Encryption using Cat Map and Chaotic Logistic Function (pp. 106-115)

Ranu Gupta (1), Rahul Pachauri (2), Ashutosh K. Singh (3)*

(1, 2) *Jaypee University of Engineering and Technology, Raghogarh, Guna (M.P.) India, 473226*

(3) *Thapar Institute of Engineering and Technology University, Patiala, 147004, India*

[Full Text: PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

15. PaperID 31101741: The Methodology of Data Collecting and the Real Time Synchronization in ETL (pp. 116-122)

Zijadin Krasniqi (1), Enver Ahmeti (2), Adriana Gjonaj (3)

(1) *Ph.D. in Information Systems, Information Technology, Pristina, Kosovo.*

(2) *M.sc in Information Technology, Pristina, Kosovo.*

(3) *Professor emeritus, European University of Tirana, Albania*

[Full Text: PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

16. PaperID 31101752: Conceptual Model of Intelligent Collaborative Educational System: Possible Solutions (pp. 123-128)

Sabina Katalnikova, Faculty of Computer Science and Information Technology, Riga Technical University, Riga, Latvia

Leonids Novickis, Faculty of Computer Science and Information Technology, Riga Technical University, Riga, Latvia

[Full Text: PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

17. PaperID 31101754: Develop a Distributed Intrusion Detection System based on Cloud Computing (pp. 129-132)

*Khoi Nguyen, Datic Lab, DaNang University, DaNang, Viet Nam
Trang Dang Thi, IT Faculty, PhamVanDong University, QuangNgai, VietNam*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

18. PaperID 31101755: A comprehensive approach on RLE and ECC (Elliptical Cure Cryptography) using Mean Square Error (MSE) feature (pp. 133-138)

*Ayushi Mathur, Department of Computer Engineering, Government Women Engineering College, Ajmer, India
Dr. Varun Prakash Saxena, Department of Computer Engineering, Government Women Engineering College, Ajmer, India*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

19. PaperID 31101757: Enhanced Multi-Modal Biometric Based Security Scheme With Feature Based Machine Learning Approach (pp. 139-143)

*Rakhi Choudhary, Er. Krishan Kumar, Dr. Himanshu Monga
Department of Computer Science Engineering, Jan Nayak Chaudhary Devi Lal Memorial College of Engineering*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

20. PaperID 31101760: Evaluating Similarity using near Set Theory for Plastic Surgery Face Images (pp. 144-149)

*Sujata G. Bhele, Electronics Engg, Priyadarshini College of Engg, Nagpur, India
Vijay H. Mankar, Electronics and Telecommunication Engg, Government Polytechnic, Nagpur, India*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

21. PaperID 31101761: Face Image Recognition Based on Linear Discernment Analysis and Cuckoo Search Optimization with SVM (pp. 150-165)

*Jamal Mustafa AL-Tuwaijari & Suhad Ibrahim Mohammed
Department of Computer Science - College of Science - University of Diyala – Iraq*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

22. PaperID 31101764: Generic Synthesis System Based on Agile Methodology for Multimedia Mobile Web Learning Modules (pp. 166-171)

*Linda lafta Gashim, Department of Computer Science, College of Science / Mustansiriyah University, Baghdad, Iraq
Assist. Prof. Dr. Karim Qasim Hussein, Department of Computer Science, College of Science / Mustansiriyah University, Baghdad, Iraq*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)

23. PaperID 31101765: Data Mining Methods for Worm Detection Using Variable Length Instruction Sequences (pp. 172-183)

Muazzam Ahmed Siddiqui & Safaa Alkatheri

Department of Information Systems, Faculty of Computing & Information Technology, King Abdulaziz University

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

24. PaperID 31101768: Tuning of Canny Image Edge Detection (pp. 184-187)

Jamil A. M. Saif,

(1) Hodeidah University, Hodeidah, Yemen

(2) University of Bisha, Bisha KSA

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

25. PaperID 31101770: The Necessity of Developing a Standard for Exchanging a Chain of Custody of Digital Evidence Data (pp. 188-191)

Jasmin Cosic, IT Section of Police Administration, MoI of Una-sana canton, 502.V.bbr, Bihać, Bosnia and Herzegovina

Miroslav Baca, Faculty of Organization and Informatics, University of Zagreb, Pavlinska 2, 42 000 Varaždin, Croatia

Petra Grd, Faculty of Organization and Informatics, University of Zagreb, Pavlinska 2, 42 000 Varaždin, Croatia

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

26. PaperID 31101772: Survey of Research Challenges in Cyber Physical Systems (pp. 192-199)

Swati Nikam, Research Scholar, DIT, Pune, India

Dr. Rajesh Ingle, Sr. Member, IEEE

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

27. PaperID 31101773: Detecting Money Laundering in a Financial System Based on Genetic Algorithm (pp. 200-208)

Ramadan Mahmood Ramo, Department of Management Information systems, University of Mosul

Prof. Dr. Khalil Ibrahim Alsaif, Department of Computer, University of Mosul

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

28. PaperID 31101778: Design of a New Small Antenna for Passive UHF RFID Tags (pp. 209-214)

Ines Frigui, Mohamed Salah Karoui, Hamadi Ghariani, Mongi Lahiani

Laboratory of Electronics and Technologies of Information (LETI)

National School of Engineers of Sfax (ENIS), Sfax, Tunisia

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

29. PaperID 31101780: Efficient Multi-Level Authentication for Cloud API based on RestPL (pp. 215-222)

*M. J. Balachandran & Dr. P. Sujatha
School of Computing Sciences, Vels University, Chennai, India*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

30. PaperID 31101781: Bitmap Indexes for Faster Query Execution (pp. 223-228)

*Samadhi P. Kumarasinghe, Nishadi D. Kirielle, Malmee Weerasinghe, Sasini Madhumali, Amal S. Perera, Sachini Herath, Amila De Silva, Lasantha Fernando
Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

31. PaperID 31101783: Optical Flow for Robot Navigation (pp. 229-237)

*Nixon Adu-Boahen, Department of Computer Science, KNUST, Kumasi, Ghana
J. B. Hayfron-Acquah, Department of Computer Science, KNUST, Kumasi, Ghana
J. K. Panford, Department of Computer Science, KNUST, Kumasi, Ghana*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

32. PaperID 31101789: Dataset of Graphs and Sub-graphs: Storage Representation and its Graphical form (pp. 238-245)

*Rachna Somkunwar (1) & Dr. Vinod M. Vaze (2)
(1) JJT University, Research scholar, Rajasthan, India
(2) JJT University, Assistant Professor, Rajasthan, India*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

33. PaperID 31101790: Fingerprint Image Retrieval using Statistical Methods (pp. 246-249)

*Sudhir Vegad, Department of Information Technology, A D Patel Institute of Technology, New V V Nagar, India
Dr. Tanmay Pawar, Electronics Department, BVM Engineering College, Vallabh Vidyanagar, India*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

34. PaperID 31101791: Proactive Detection of Catastrophe Trends for Rescheduling Real-Time Systems with Scenario Shift (pp. 250-268)

*A. Christy Persya (1), T. R. Gopalakrishnan Nair (2)
(1) Advanced Real-Time Computing Group, RRGI Research Centre, Bangalore, India.
(2) Advanced Real-Time Computing Group, RRGI Research Centre, Rector RRGI & Visiting Professor, NIAS, Bangalore, India.*

Full Text: [PDF](#) | [Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

35. PaperID 31101792: Big Data Analytics Framework for Peer-To-Peer Botnet Detection Using Random Forest and Deep Learning (pp. 269-277)

*Saraladevi D., PG Student, Department of Computer Science and Engineering, Pondicherry Engineering College
Sathiyamurthy K., Associate Professor, Department of Computer Science and Engineering, Pondicherry*

Engineering College

*Vijayaprabakaran K., PhD Scholar, Department of Computer Science and Engineering, Pondicherry Engineering
College*

Full Text: PDF [[Academia.edu](#) | [Scopus](#) | [Scribd](#) | [Archive](#) | [ProQuest](#)]

Extended Multi Queue Job Scheduling in Cloud

Neeraj Kumar Pandey

Research Scholar, Uttarakhand University, Dehradun Uttarakhand

Email: dun.neeraj@gmail.com

Sumit Chaudhary

Research Scholar, Uttarakhand University, Dehradun Uttarakhand

Email:iimtsumit@gmail.com

N.K.Joshi

Director, Uttarakhand Institute of Technology, Uttarakhand University,

Dehradun Uttarakhand

Email: nkjoshi2001@yahoo.com

Abstract:

Cloud computing refers to a ‘Utility Computing’ where the services can be used as per the requirements and the demand can be expanded and contracted as per the business needs. One of the major features provides the facility to pay as per the use. Scheduling is a major challenge in any cloud environment where multiple tasks are arriving at same time and CSP has to manage them pretending the unlimited resources for any request. While cloud service provider have limited amount of resources to fulfill all request. Min min algorithm used to minimize the make span time of task execution. Each algorithm is designed in such a way that by using it CSP must achieve user satisfaction. In this paper an improved economy based scheduling strategy is proposed after analyzing traditional algorithm which is based on task length, priority and cost of execution. These parameters are considered before the allocation of the resources so this strategy shows the considerable improvement in the existing strategies.

Keywords: Cloud Service Provider (CSP), Resource Allocation Strategy, Cloud computing, Task priority, Task length.

Introduction: A cloud is a technology where many users are connected to many systems through a network. A scalability property allows the user to demand any service at any time without bothering about the availability of resource to fulfill the corresponding request. So this situation creates a great challenge of managing the resources so before the CSP in such a way so that the entire request must be obeyed within time constraints.

Resource Scheduling is one of the major challenges in cloud computing. There are numerous strategies used by the CSP depending upon the situation or demand of the user task. When a task is assigned to execute the data center will take care of how it will execute. The following figure will illustrate the basic architecture of assigning any task to data center. The tasks are submitted by different users to the datacenter broker. The datacenter broker acts like a mediator between the user and data center.

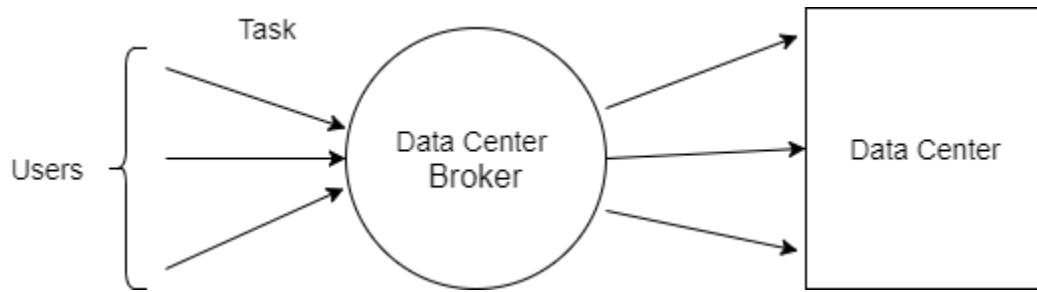


Figure.1

Min min strategy considers the task length as the level for executing the assignments. An ideal condition happens when undertaking with least length is executed first. With a specific end goal to accomplish this ideal condition datacenter intermediary ought to consider this factor. But on the other hand the order of execution of the task may differ if the other parameter like priority and the cost of the demanded resource to execute the task is considered. The main idea about the proposed approach is to consider the both criteria i.e. priority of the task and resource cost to execute the task.

2. Related work:

A lot of resource scheduling algorithm is available and used nowadays. The selection of scheduling algorithm depends on the parameters it works. But the performance is always an issue to analyze. Three parameters are considered in the proposed approach i.e size of the task, priority of the task and cost of execution of task. Many researchers are working with the study area multi queue scheduling algorithm [2], OCRP [3], Activity based costing [4][5], Simulated Annealing (SA) [6]. A lot of survey [7][8] have been done to identify the characteristic of different scheduling strategies so that the optimized scheduling method can be found.

2.1 Priority based: priority is a major concern in job scheduling algorithms. There are numerous algorithms which performs job scheduling by considering priority as a major factor. Analytical hierarchy process(AHP)[9] uses multi criteria decision making and multi-attribute decision making models to evaluate the priority of the task.

2.2 Cost Based Resource Scheduling: this strategy is based on the cost of the allocation of resource to any request received by the CSP. CSP calculates the cost by considering the resource cost and the performance of the requested resource cost. If the execution time is increases by allocating the resource then it will increase the execution cost which will be combined with the resource cost. Then finally the cost of allocation of resource gets increased. Paper [10] calculates the cost by reservation and demand strategy in which resources are reserved as per the demand generated in the future.CSP can change the resource before it is consumed by the user.

2.3 Energy efficiency Based: A decentralized architecture of energy aware resource management is proposed in the paper [15]. A problem is defined to minimize the energy consumption of resource allocation policy in the cloud data center. It will produce the environment friendly strategy to manage the energy efficient data center development concept.

2.4 Non Preemption Based: to implement the non-preemptive (Parallel) Scheduling Naphele scheduler has been introduced in the paper [12]. Naphele calculated the critical time (maximum time allowed to take by any process) so that if the execution reaches the critical time it will be suspended and added to the waiting queue. Naphele [11] provided a framework for the parallel data processing in the cloud which completely exploits the dynamic resource allocation.

2.5 Heuristic Based: in this paper [13] a heuristic based cost optimization algorithm is proposed. Heuristic provides optimal solution of resource allocation problem under some restriction. Although resource provisioning is a NP hard problem but heuristic based solution provides optimum result in various conditions.

2.6 Reliability Based: Minimum failure resource allocation (MFRA)[14] problem have been discussed for requested VMs residing in the multiple data centers located in different geographical locations. MFRA is a NP complete problem but ILP framework is proposed to provide solution of small scale problem of distributed cloud. Two heuristic algorithms are proposed as a large scale topology which gives a optimal result in a small scale topology.

2.7 Replication Based: in this strategy cloud is divided into sub cloud and submitted job is replicated to all sub cloud following different resource provision algorithm. Each sub cloud executed the same replicated job. If any sub cloud finishes the job it will broadcast the finish message to all other sub cloud

to stop execution. It is not a feasible solution of resource provisioning because of wastage of time and resource use in the partial execution of the job by sub cloud.

3. Proposed Approach

Jobs having equal priority or no economic urgency will be divided into two queues based on burst time. These queues will execute the jobs from both queues one by one as seen in figure 2[16][8]. But if the jobs come with different priority then queue selector will put them into a priority queue which will execute the jobs according their priority (preemptive method). Dynamic job selector will select the job dynamically between the two different types of queue for the cloud environment. The order of execution of jobs is:

1) For equal priority jobs

There are two types of queue one for smaller burst time i.e.70% of total job and other is large burst time i.e.30% of total jobs. So the order of execution is

P1→P2→P3→P4→P5→P6→P7→P8

2) For different priority jobs:

P2→p1→p3→p4→p8→p7→p5→p6

Pseudo code for priority scheduling:

1. For all T_s (Task Submitted)
2. Find the priority of submitted task.
3. Maintain the ready queue based on newly arrived job
4. For all newly arrived job
 - If (priority of new job $T_n >$ Priority of job in execution T_e)

$$T_e = T_n$$
 - 5. Run the job queue using priority scheduling
 - 6. Else
 - 7. Divide the jobs according their burst time and use traditional Scheduling

Input Data

Job ID	P1	P2	P3	P4	P5	P6	P7	P8
Burst Time(ms)	15	88	85	42	35	20	92	72

Table.1(FCFS)

Job ID	P1	P6	P5	P4	P8	P3	P2	P7
Burst Time(ms)	15	20	35	42	72	85	88	92
Table.2 (SJF)								
Job ID	P1	P4	P5	P2	P3	P6	P8	P7
Burst Time(ms)	15	42	35	88	85	20	72	92

Table.3 (CBA)

Job ID	P1	P8	P6	P3	P5	P2	P4	P7
Burst Time(ms)	15	72	20	85	35	88	42	72
Table.4 (MQS)								

The priority based queue will give result based on their priority. The order of execution will be P1→P8→P6→P3→P5→P2→P4→P7

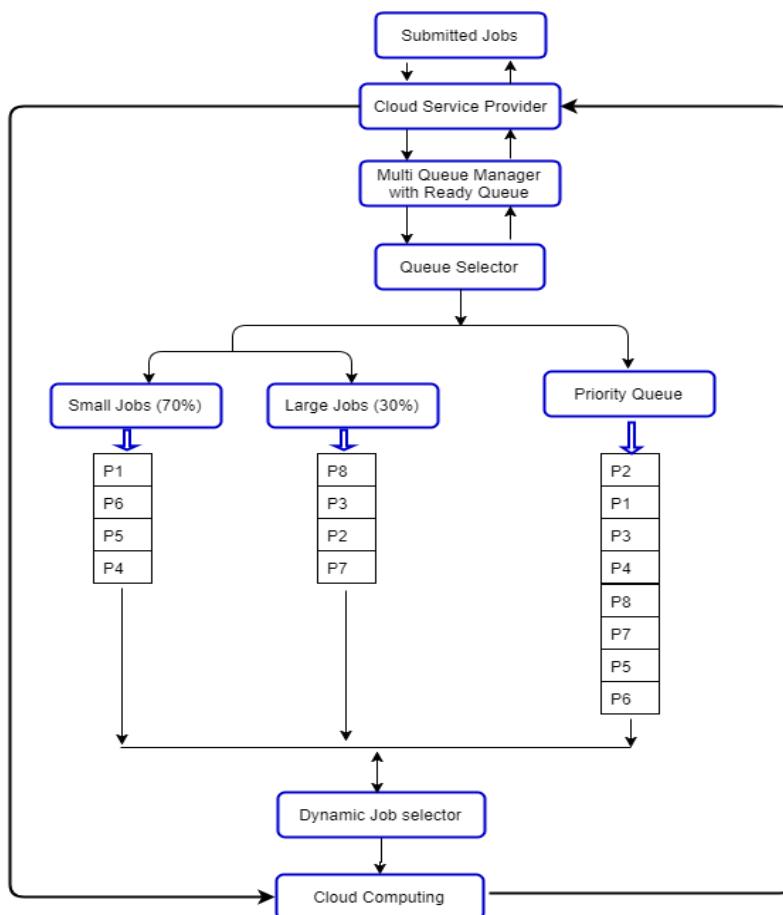
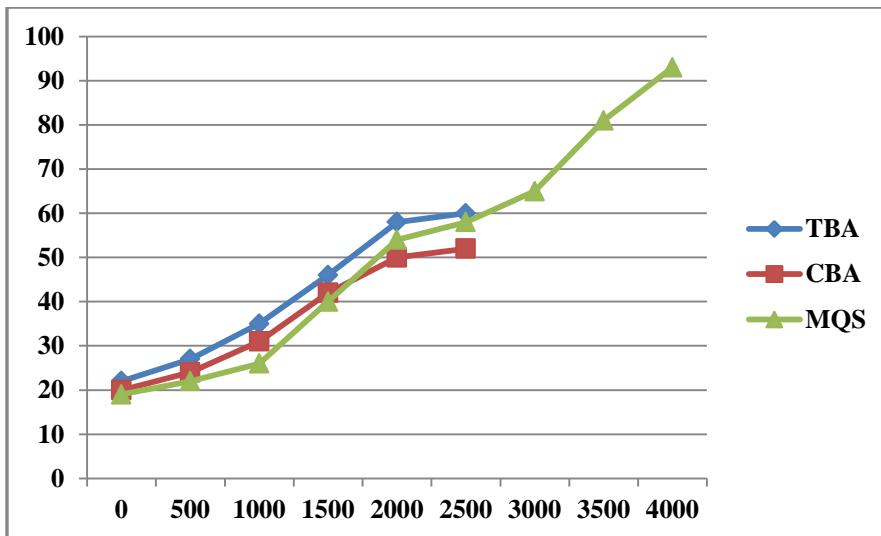


Figure.2

Result and Discussion

The result of proposed multi queue scheduling strategy gives optimized result when compared to some well-known job scheduling strategy. The proposed approach divides the job into two parts based on their priority i.e. multi queue for equal priority queue for different priority process. It gives better result in real time systems. The proposed strategy provides optimum result for cloud environment and enhances the resource utilization in cloud environment. Cloud simulator is used to implement the proposed strategy which simulates the allocation of processing element (PE) which is shared to the cloudlet job. Cloudlet attains the PE depends on the number of jobs in the ready queue. When the finishes the execution it is removed from the queue (Execution Set). Cloudsim schedules all jobs depend on the current position of the job present in ready queue. The simulation of MQS and cloudlet execution into small queue P1, P6, P5, P4 and P8, P3, P2, P7 are queued. The scheduler takes one job from queue in every iteration for equal priority jobs and if priority is different the execution is maintained in different queue in which priority of job is calculated first. If any high priority jobs arrive then current executing job is preempted. It follows the priority with preemption policy for current execution set. It provides optimized result in cloud environment.



Graph 1: comparison of the proposed method with existing method

Conclusion and future scope

The scheduling methodology simply defining the efficient way of execution of job according the demand of the user. If users give some priority to the job then it is handled by priority scheduler but if it does not have any priority then it is considered as zero priority and put into a queue as per the burst time which is handled by traditional scheduler. The future of this method can add economy based queue after calculating the cost of the execution.

Reference:

- [1] Antony Thomas, Krishnalal G, Jagathy Raj V P “Credit Based Scheduling Algorithm in Cloud Computing Environment” International Conference on Information and Communication Technologies (ICICT 2014) pp. 913-920.[2] AV.Karthick, Dr.E.Ramaraj, R.Ganapathy Subramanian “An efficient multi queue job scheduling for cloud computing” IEEE World Congress on Computing and Communication Technologies 2014 pp.164-166.
- [3] Abhishek Vichare, Zenia P. Gomes, Noella Fernandes and Flavin Cardoza “Cloud computing using OCRP and virtual machines for dynamic allocation of resources” IEEE International Conference on Technologies for Sustainable Development (ICTSD-2015), Feb. 04 – 06, 2015, Mumbai, India.
- [4] Qi cao,Zhi-bo Wei,Wen-mao Gong. An Optimized Algorithm for Task Scheduling Based On Activity Based Costing in Cloud Computing.
- [5] J Blythe, S Jain, E Deelman, Y Gil, K Vahi. Task scheduling strategies for workflow-based applications in grids. Cluster Computing and the Grid; 2005.
- [6] Diptangshu Pandit, Samiran Chattopadhyay, Matangini Chattopadhyay “Resource allocation in cloud using simulated annealing” IEEE Applications and Innovations in Mobile Computing (AIMoC) 2014).
- [7] T. a1, A. Jaya Lakshmi 2, Vuyyuru K. Reddy “Resource Allocation Methods in Cloud Computing : Survey”international journal on emerging trend in technology ISSN: 2350 – 0808 September 2015 Volume 2 Issue 2 pp 416-419.
- [8] Neeraj Kumar pandey, Sumit Chaudhary,N.K.Joshi “Resource allocation strategies used in cloud computing: A critical analysis”IEEE 2nd International Conference on Communication Control and Intelligent Systems (CCIS), 2016.
- [9] Swachil J. Patel, Upendra R. Bhoi “Improved priority based job scheduling algorithm in cloud computing using iterative method” IEEE Fourth International Conference on Advances in Computing and Communications 2014 pp.199-202.
- [10] Abhishek Vichare, Zenia P. Gomes, Noella Fernandes and Flavin Cardoza “Cloud computing using OCRP and virtual machines for dynamic allocation of resources” IEEE International Conference on Technologies for Sustainable Development (ICTSD-2015), Feb. 04 – 06, 2015, Mumbai, India.
- [11] Warneke, Daniel, and Odej Kao. "Nephele: efficient parallel data processing in the cloud." Proceedings of the 2nd workshop on many-task computing on grids and supercomputers. ACM, 2009.
- [12] Gandhali Upadhye,Truspti Dange “Cloud resource allocation as nonpreemptive approach” IEEE 2nd International Conference on Current Trends in Engineering and Technology(ICCTET)2014 pp.352-356.

- [13] Sunirmal Khatua, Preetam Kumar Sur, Rajib Kumar Das, and Nandini Mukherjee “Heuristic-Based Resource Reservation Strategies for Public Cloud” IEEE transactions on cloud computing, vol. 4, no. 4, october-december 2016.
- [14] Yi Zhu, Yan Liang, Qiong Zhang, Xi Wang, Paparao Palacharla and otoyoshi Sekiya “Reliable Resource Allocation with Weighted SRGs for Optically Interconnected Clouds” IEEE Global Communications Conference 2014 pp. 2186-2191.
- [15] Anton Beloglazov and Rajkumar Buyya “Energy Efficient Resource Management in Virtualized Cloud Data Centers” 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing 2010 pp. 826-831.
- [16] AV.Karthick, Dr.E.Ramaraj, R.Ganapathy Subramanian “An efficient multi queue job scheduling for cloud computing” IEEE World Congress on Computing and Communication Technologies 2014 pp.164-166.

Analyzing and Predicting Hash tags Popularity on Twitter

Krishna Vasu Katarla

M.Tech Final Year, Dept of CSE,

Sri Vasavi Engineering College

Pedatadepalli, Tadepalligudem, A.P, India

vasu.krish2@gmail.com

Dr. Vedula Venkateswara Rao

Associate Professor, Dept of CSE,

Sri Vasavi Engineering College

Pedatadepalli, Tadepalligudem, A.P, India

Venkatvedula2017@gmail.com

Abstract— Twitter and micro blogging services have became in dispensable sources of information on today's web. Understanding the main factors that make certain pieces of information spread quickly on these platforms can be deceiving for the analysis of opinion formation. Twitter Data is one of the large amounts of sized data, because it is having a millions of tweets every day. It is one of the largest social media site. We are using this twitter data for the business purpose and industrial or social purpose according to our data requirement and processing the data. It is very large amount of sized data increasing every second that is known as big data. In today's world twitter is not only a social networks but also an increasingly interesting news important information broadcasting mechanisms and new media's that have many followers are not important information sources. Thus becomes very important to understand news related propagation from news media called as super nodes. A useful practice in social media network analysis is to predict future popularity of event good platform to perform such analysis with twitters topic structure on mind, the problems can be stated as : "knowing current and previous tweet activity in future or can we predict if it will become more popular and if so how much. This understanding prediction benefits many applications such as social management advertisement and interaction optimization between media and followers.

Keywords- micro blog, twitter, retreat, hash tag, super node, registration analysis, predictions

I. INTRODUCTION

Twitter, with its public discussion model, is a good platform to predict future popularity of a topic or event. Knowing current and previous tweet activity for a hash-tag (#), we can predict if it became more prominent and trendy in the future and if yes by how much. In the times of information age, the magnitude of online social media activity has reached unprecedented level. Hundreds of millions of users spend hours online everyday to stay connected and communicate with the rest of world. Millions of users participate in these social networks of Social awareness streams. [19] People generate huge amount of data everyday on various social media networks, which in aggregate indicate the interests and current attention of the local and global communities. There are many events and topics discussed on Twitter. Some topics may get a lot attention and some may not. Some of these topics become very popular and focus of interests for large number of people. The connections and the nature of social network let

information disseminate to large number of other people, a phenomena known as going "viral". These popular topics of discussions are also called "trends" in the social network. These trends are very dynamic and temporal in nature which exposes the expose the aggregate interests and attention of global and local communities. Trends in social networks are of high significance and a major point of interest in both the industry [19] and the research community [8, 15]. Many applications on web and business can be immensely benefitted from knowing what is currently \trending", which represents an answer to the age-old query what are people talking about? [9]. From stock exchange making real time decision to search engines delivering more updated, relevant search results.

Figures 1.1, 1.2 Twitter is one of the most popular social networking and micro-blogging service, which had more than 200 Million registered users by 2013, producing 400 Millions tweets everyday [17].

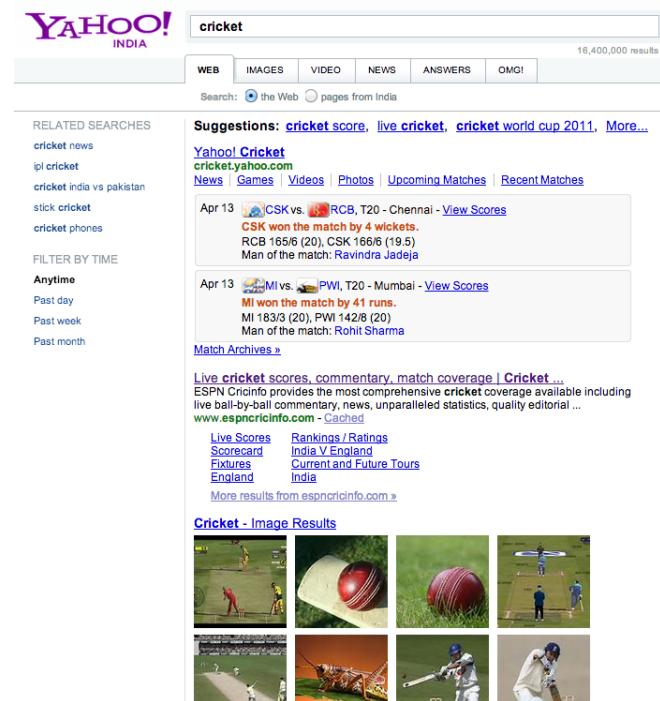


Fig 1.1 Trending topic \Cricket"

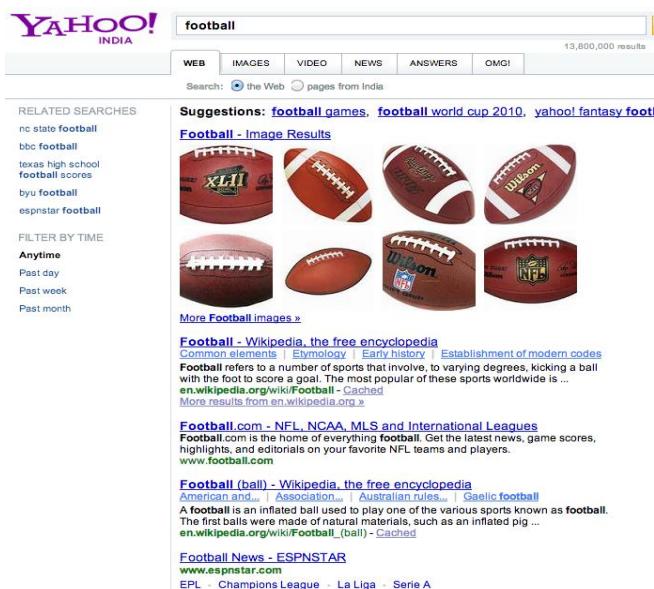


Fig 1.2 Non-trending topic "Football"

A. Motivation

In twitter more than 19% of the tweets are about organizations or product brands, less than 20% of which are shown to have significant sentiment. Predicting the tweets which are likely to stimulate users' interests can improve the sale and marketing of different products and brands. Online advertisements could use such predicted messages to efficiently target the locations of networks which are visited the most. Moreover successful predictions can also increase user satisfaction by providing them with more attractive contents. Media companies could learn how to effectively generate buzzes for new films and shows. In political campaigning, groups could learn who they should target in order to successfully spread their message. Predicting the popularity of content in twitter is also quite important for several other purposes such as viral marketing, popular news detection, personalized message recommendation and trend analysis. Users with many connections can suffer from information overload. It is quite important to filter information flow for the end users and to provide them with important tweets. Popularity prediction is also helpful in personalizing the content and finding the right tweets for end users. On the other hand, understanding how and why a tweet becomes popular, can help to gain a better insight into how the information is dispersed over the network. In the case of marketing, predicting popular tweets is quite useful for determining what are the trending topics and products.

In this work we developed an automatic learning-based approach to predict the popularity of content in tweeter. Automatic prediction with machine power has much lower costs compared to human-based work

The problem of popularity prediction in twitter has been studied in some previous works. In most of the recent works the popularity of tweets is defined as the number of retweets

since retweeting is potentially the most effective way to disseminate messages due to its viral nature.

B. Problem Statement

Our problem statement can be formally defined as follows. Create an automated system, which takes in continuous stream of raw tweets, processes these tweets to filter out 'noise' and get relevant informative tweets. Further mine these 'filtered' tweets to detect and predict evolving 'trending' topics in their very early stage.

There are three key aspects to this problem:

- Creating a model to process real time tweets feed to filter out the 'noise' and output only relevant tweets.
- Creating a system to store the social graph of twitter users such that it can be efficiently used as a data structure for answering various user queries and get relations for a given user.
- Creating a model to distinguish a 'trending' topic from 'non-trending' to make

II. RELATED WORK

A. Introduction

The problem of popularity prediction in social networks has always been widely studied. This problem has not only been studied in conjunction with twitter, but also in connection with other social networks. In this section we provide an overview of the existing approaches to popularity prediction in social networks, we discuss the related work and we elaborate on the advantages and limitations of existing methods.

B. Information Dissemination in Social Networks

A growing line of research has been followed on information dissemination through social networks. These studies propose that network cascades can play an important role as mediums for the dissemination of various information. These studies tend to be based on the idea that the information is spread by various infection mechanisms Under the same category, Kempe et al. (2003) studied a combinatorial optimization problem sometimes known as the influence maximization problem. The problem involves finding a small set of seed nodes in social networks to target initial activation so that the largest expected spread of information can be yielded. However, the exact computation of information cascades is an NP-hard problem (Chen et al., 2010). Information diffusion has been studied in several online social networks, such as Flicker.

C. Predicting the Popularity of Content in Social Networks

Due to the advent of web 2.0, user-generated content has increased dramatically. There are various types of contents that can be generated by users, such as comments and reviews on photos, movies and products. Most of these web 2.0 services connect the user with other users through social network, thus producing a social graph. For instance, in micro blogging services such as Twitter this social graph is called a follower network. Any content generated from a user becomes

visible to all of his/her followers and each of these contents has the chance to be re-posted by these followers who subsequently disperse the content over the social network. Re-posting, commonly known as retweeting, gives post the chance to become popular. The problem of popularity prediction in social networks has been widely studied.

D. Popularity Prediction in Twitter

Due to the popularity of the twitter micro blogging service there have been many studies on twitter. A great amount of work has been done to predict the popularity of tweets in this network. In this section we first justify why users do retweeting in twitter by reviewing the related literature and then we briefly explain the related work on popularity prediction on twitter social network. Understanding how users tweet and their motivations for tweeting is potentially important for predicting whether a tweet will be popular or not. In fact discovering what contents users choose to retweet can help to explain why a particular tweet becomes popular. The motivations for the act of retweeting are well explored in the study done by Boyd et al. (2010). They highlighted the main reasons for retweeting as given by users. They introduced 10 different motivations for retweeting such as commenting on tweets, propagating tweets to new audiences, to inform specific persons or groups and to save tweets for future personal access.

E. Popularity Predictions and Recommendations

The problems of popularity predictions and recommendations are similar in some aspects. Both problems try to identify influential contents. While in popularity prediction problems the focus is more on the popularity of content, in recommender systems the focus is more on the user, the goal being to recommend the items to a user which satisfy him the most. Predicting the popularity of contents can be quite useful in connection with making recommendations.

F. Social Network Prediction Applications

Predictive models analyze past information to assess how likely it is than an event will occur in the future. Although human experts could have greater accuracy they are not scalable and do not work properly in cases when events have very low or high probability and they are definitely more expensive compared to the computer-based approach.

III. DATA COLLECTION – DATA SET

Twitter data is collected by querying popular hash-tags related to the 2015 Super Bowl spanning a period starting from 2 weeks before the game to a week after the game. We use this data to train a regression model and then use the model to make predictions for other hash-tags. The test data consists of tweets containing a hash-tag in a specified time window, and we have then used our model to predict number of tweets containing the hash-tag posted within one hour immediately following the given time window. Twitter is a social networking and micro blogging service that allows users to post real time messages, called tweets.

Tweets are short messages, restricted to 140 characters in length. Due to the nature of this micro blogging service, people use acronyms; make spelling mistakes, use emoticons and other characters that express special meanings. Following is a brief terminology associated with tweets.

- Emoticons: These are facial expressions: pictorially represented using punctuation and letters; they express the user's mood
- Target: Users of Twitter use the @ symbol to refer to other users on the micro blog. Referring to other users in this manner automatically alerts them.
- Hash tags: Users usually use hash tags to mark topics. This is primarily done to increase the visibility of their tweets.

IV. PREPROCESSING

A. Tokenization

After downloading the tweets using the tweet id's provided in the dataset, we first tokenize the tweets. This is done using the Tweet-NLP developed by ARK Social Media Search. This tool tokenizes the tweet and returns the POS tags of the tweet along with the confidence score. It is important to note that this is a twitter specific tagger in the sense it tags the twitter specific entries like Emoticons, Hash tag and Mentions too. After obtaining the tokenized and tagged tweet we move to the next step of preprocessing.

B. Remove Non-English Tweets

Twitter allows more than 60 languages. However, in our work currently focuses on English tweets only.

C. Replacing Emoticons play an important role in determining the sentiment of the tweet. Hence we replace the emoticons by their sentiment polarity by looking up in the Emoticon Dictionary.

D. Remove Url

The url's which are present in the tweet are shortened using TinyUrl due to the limitation on the tweet text. These shortened url's did not carry much information regarding the sentiment of the tweet. Thus these are removed.

E. Remove Target

The target mentions in a tweet done using '@' are usually the twitter handle of people or organization. The information is also not needed to determine the sentiment of the tweet. Hence they are removed.

F. Replace Negative Mentions

Tweets consist of various notions of negation. In general, words ending with 'nt' are appended with a not. Before we remove the stop words 'not' is replaced by the word 'negation'. Negation play a very important role in determining the sentiment of the tweet. This is discussed later in detail.

G. Hash tags

Hash tags are basically summarizer of the tweet and hence are very critical. In order to capture the relevant information

from hash tags, all special characters and punctuations are removed before using it as a feature.

H. Sequence of Repeated Characters

Twitter provides a platform for users to express their opinion in an informal way. Tweets are written in random form, without any focus given to correct structure and spelling. Spell correction is an important part in sentiment analysis of user-generated content. People use words like 'cooooo' and 'hunnnnnngry' in order to emphasise the emotion. In order to capture such expressions, we replace the sequence of more than three similar characters by three characters. For example, wooooow is replaced by wooow. We replace by three characters so as to distinguish words like 'cool' and 'coooooool'.

I. Numbers

Numbers are of no use when measuring sentiment. Thus, numbers which are obtained as tokenized unit from the tokenizer are removed in order to refine the tweet content.

J. Nouns and Prepositions: Given a tweet token, we identify the word as a Noun word by looking at its part of speech tag given by the tokenizer. If the majority sense (most commonly used sense) of that word is Noun, we discard the word. Noun words don't carry sentiment and thus are of no use in our experiments.

K. Stop-word Removal

Stop words play a negative role in the task of sentiment classification. Stop words occur in both positive and negative training set, thus adding more ambiguity in the model formation. And also, stop words don't carry any sentiment information and thus are of no use to us. We create a list of stop words like he, she, at, on, a, the, etc. and ignore them while scoring the sentiment.

V. ARCHETECTURE

In this section we build feature base model to perform classification using a combination of both these models. Our approach can be divided into various steps. Each of these steps are independent of the other but important at the same time. Figure 1 represent the approach for training and explains in detail about different steps in training process.

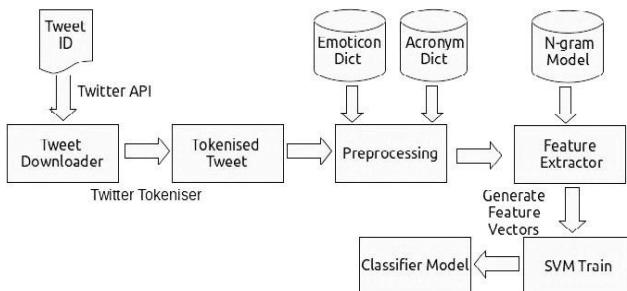


Figure 4.1 Flow Diagram of Training Data (A Hybrid Model)

Figure 4.2 represents testing the model where it describes various steps for testing the application with varying data sets.

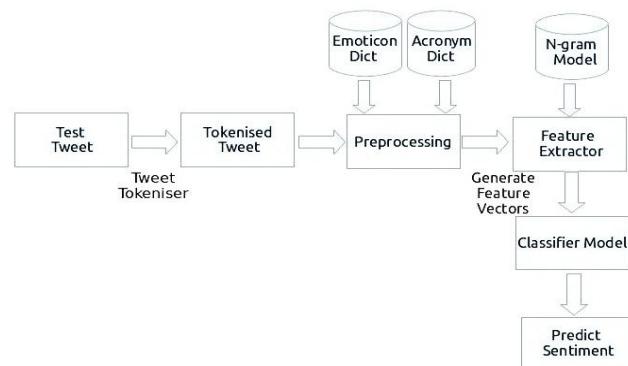


Figure 4.2 Flow diagram of testing (A Hybrid Model)

VI. MODEL TO MINE TWEETS

A. Linear Regression

A linear regression model was created using 5 features to predict number of tweets in the next hour, with features extracted from the tweet data in the previous hour. The features used to create the model were,

1. Numbers of Tweets (Class Variable)
2. Total Number of Re-tweets - (metrics/citations/total)
3. Sum of the number of followers of the users - (authors/followers)
4. Maximum number of followers of the users posting the hash tag
5. Time of the data - Obtained using the post-time of the tweet

The same hour-window approach was employed to calculate all the features. The output variable for each hour-window was the tweet count for the next hour-window. The model was trained using the OLS stats model library.

B. Regression Model with Extra Features

A new regression model was created using custom extra features (including original features considered in Question 2.) that were considered based on various papers and observation of the data. The new features considered were as follows,

1. Numbers of Tweets (Class Variable)
2. Total Number of Re-tweets - (metrics/citations/total)
3. Sum of the number of followers of the users - (authors/followers)
4. Maximum number of followers of the users posting the hash tag
5. Time of the data - Obtained using the post-time of the tweet
6. Ranking Score - (metrics/ranking_score)
7. Impression Count - (metrics/impresion) - Measures the number of times a user is served a Promoted

Tweet either in time-line or on search

8. Favorite Count - (tweet/favorite_count) - Number of tweets favorite's by users
9. Number of Users per hour - (tweet/user/id) - Counted number of users posting per hour
10. Number of Long Tweets per hour - (title) - Counted the number of tweets with length > 100 characters.

C. Cross Validation

This function gives internal and cross-validation measures of predictive accuracy for ordinary linear regression. The data are randomly assigned to a number of 'folds'. Each fold is removed, in turn, while the remaining data is used to re-fit the regression model and to predict at the deleted observations. Cross validation is an approach to partition the sample data into a training (or model-building) set, which we can use to develop the model, and a validation (or prediction) set, which is used to evaluate the predictive ability of the model.

VII. EXPERIMENTAL RESULTS AND ANALYSIS

A. Tweet Data Statistics

The training tweet data was loaded and statistics for each hash-tag was calculated in this question. In order to keep track of the hour count we have employed a hour-window approach. Since the tweets are all in sorted order of their posting time (firstpost_date). The first tweet is considered and the 1st hour-window is created using the formula

$$\text{end_time} = \text{start_time} + 3600 (1)$$

We loop through each tweet in the file and compare the post-time of the tweet with the end time of the present window. If it lies within the window we increase the hour-count if it doesn't we create a new window by using eqn:(1) and adding 3600 again to the end-time. At the same time a count is kept for the number of followers of users (author/followers) and the number of re-tweets (metrics/citations/total) for each tweet. The statistics calculated using the above procedure is listed below.

Table 7.1: Statistics for Each Hash tag

Hash tag	Total Tweets	Avg. # Tweets/hr	Avg. # of Followers of Users	Avg. # of Retweets
#gohawks	188135	193.5438	1596.443	2.0146
#gopatriots	26231	38.3832	1292.2031	1.4001
#gopatriots	259019	279.5503	4394.2539	1.5385
#patriots	489710	499.4200	1607.4407	1.7828
#sb49	826905	1419.886	2229.694	2.5111
#superbowl	1348766	1401.2445	3675.3394	2.3882

Analysis of the Statastics

1. Most Tweeted Hash tags per hour: #sb49 and #superbowl
2. Most Followers of Users for Hash tag : #nfl and #superbowl
3. All of the tweet data collected comprise of tweets that are not re-tweeted or are re-tweeted by very few users hence making the average re-tweet count _ 2.

In order to visualize the number of tweets in an hour a histogram was plotted for #SuperBowl and #NFL. A steep-rise can be seen for both the graphs at the same time which indicates the hour of the event.

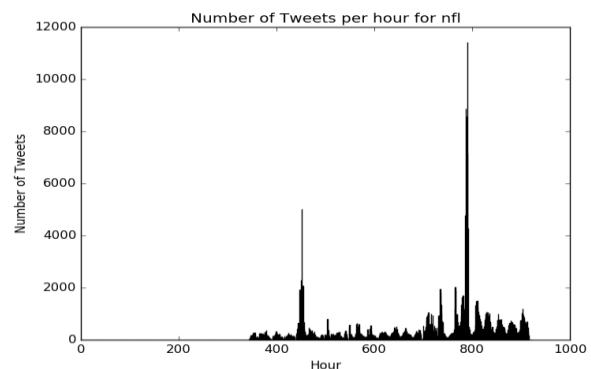
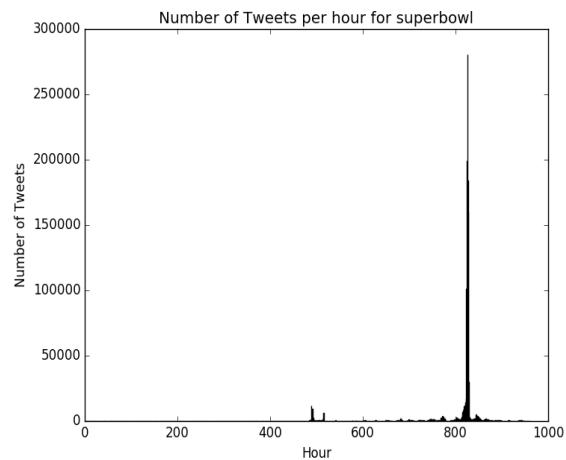


Figure 7.1: Number of tweets in hour : #NFL

Figure 7.2: Number of tweets in hour : #SuperBowl



B. Linear Regression

The same hour-window approach was employed to calculate all the features. The output variable for each hour-window was the tweet count for the next hour-window. The model was trained using the OLS stats model library. The results obtained for each of the hash tag are as follows,

Table 7.2: Model Accuracy for each Hash tag

Hash tag	Accuracy
#gohawks	41.78
#gopatriots	43.15
#gopatriots	54.69
#patriots	43.72
#sb49	58.54
#superbowl	66.13

The (p-value ¹, t-value ²) for each attribute was recorded as well, the results are as follows

Table 7.3: p-value & t-value for Model Parameters

Hash tag	#of Retweets	Σ of # of followers of users	Max. # of followers	Time of the data
#gohawks	(2.115*10 ⁵ , 4.273)	(1.066*10 ⁷ , 5,355)	(1.290*10 ⁻⁶ , -4.871)	(4.230*10 ⁻³ , 2,867)
#gopatriots	(8.732*10 ²⁷ , 11,247)	(3.517*10 ¹⁶ , -8,386)	(1.048*10 ⁻¹² , 7,278)	(8.643*10 ⁻¹ , -0.170)
#nfl	(3.525*10 ¹⁶ , 8,266)	(7,424*10 ² , 1,787)	(4.185*10 ⁻¹ , -0.809)	(2.314*10 ⁵ , 4,235)
#patriots	(4.432*10 ⁶³ , 18,053)	(6,807*10 ¹⁴ , -7,602)	(2.209*10 ⁻⁵ , 4,263)	(4.776*10 ⁻³ , 0,710)
#sb49	(4.681*10 ⁵⁶ , 17,647)	(2,329*10 ³¹ , -12,347)	(1.320*10 ⁻¹⁵ , 8,222)	(6,402*10 ⁻² , -1,855)
#superbowl	(4.897*10 ¹⁴⁹ , 31,256)	(1,612*10 ⁷ , -26,502)	(4,853*10 ⁵² , -6,145)	(7,136*10 ⁻² , 1,805)

Analysis of Results

➤ According to the definition of p-value and t-value it can be seen that the most contributing feature towards the regression model in all hash-tag files is the Number of Retweets posting a hash-tag.

➤ A fairly low accuracy is obtained for most of the hash-tag. This can be attributed to the window-size of one-hour as in the initial hours the average number of tweets are pretty low and creating a model for these sparse features is more difficult.

C. Regression Model with Extra Features

A total of 9 features were used to create the new regression model and after employing the same methodology of Question-2, features were collected using one-hour window. Since the last hour window cannot predict a tweet-count value it has been removed while creating the model. The model was tested and the results obtained were as follows

Hash tag	Accuracy
----------	----------

#gohawks	78.384
#gopatriots	53.118
#nfl	nfl 64.840
#patriots	58.793
#sb49	70.623
#superbowl	77.089

Table 7.4: Model Accuracy for each Hash-tag

As seen from the above results we have a significant increase in the accuracy of the model for each of the hash-tag, this can be attributed to features that are not sparse and have a well defined distribution through-out the period of the SuperBowl. Metrics employed in the tweet-data have been used to model the importance of the tweet for a given window frame thereby increasing the accuracy. In order to better visualize the contribution of the features in the model a scatter plot was created of the Top 3 features for each hash-tag. Since the initial hours have less number of tweets, all of the graphs exhibit clustering of values near low of tweets/hour.

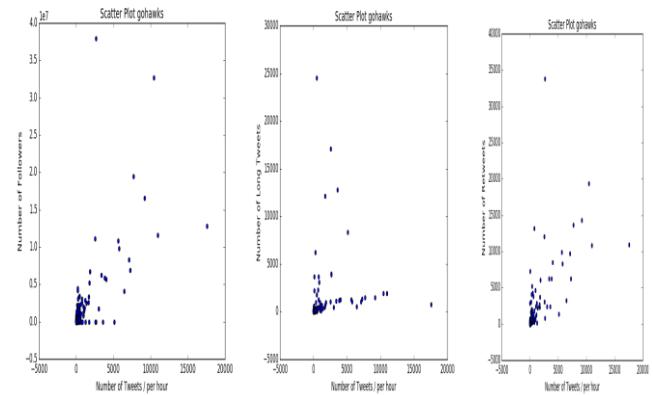


Figure 7.3: Top 3 feature for #gohawks (# of Followers, # of Retweets, # of Long Tweets)

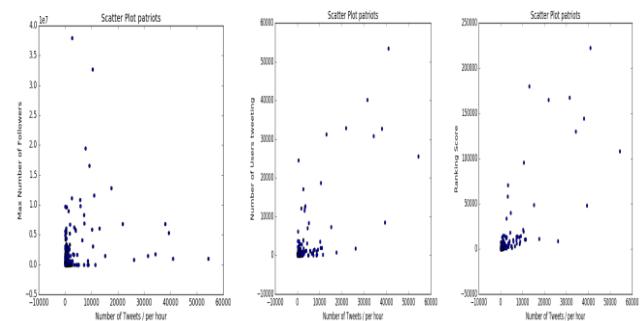


Figure 7.4: Top 3 feature for #gohawks (Max # of Followers, # of Users Tweeting, Ranking Score)

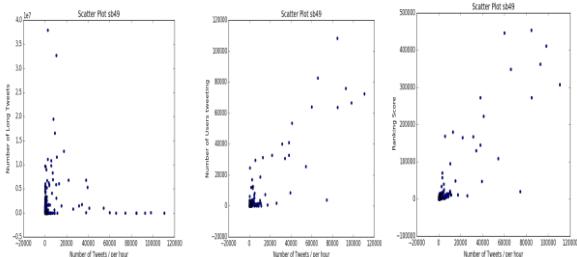


Figure 7.5: Top 3 feature for #gohawks (# of Long Tweets, # of Users Tweeting, Ranking Score)

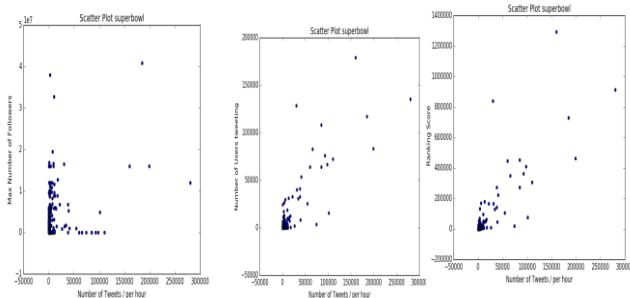


Figure 7.6: Top 3 feature for #gohawks (Max # of Followers, # of Users Tweeting, Ranking Score)

Analysis of Results

Table 7.5: Analysis of Top 3 Features Scatter Plot

Hash tag	Analysis
#gohawks	A linear proportionality can be seen in the scatter plots signifying a good relationship between all the 3 features
#gopatriots	Almost identical scatter plots with clustering towards the region of the origin
#nfl	A constant relationship can be seen for features (Favorite Count, # of Long Tweets) while a linear relationship is visible for # of Retweets
#patriots	Constant relationship for Max # of Followers feature while other two show linear proportionality
#sb49	Similar analysis to #patriots
#superbowl	Clustered regions with a very small linear deviation. Large number of instances fits a better regression model hence the higher accuracy

D. Cross Validation

This also requires the usage of same features used in Linear Regression with extra features to perform 10-fold Cross Validation across data. The accuracy results obtained across various hash-tags and over every fold given below.

Table 7.6: Average Error of 10 Fold Cross Validation

Fold Number	#gopatriots	#gohawks	#nfl	#patriots	#sb49	#superbowl
(1)	7.782	20.127	23.921	180.855	31.417	229.980
(2)	8.438	46.514	1.376	84.489	61.089	255.881
(3)	10.145	4.814	3.181	31.927	99.079	337.870
(4)	204.985	2.245	28.109	52.189	64.583	397.136
(5)	15.497	117.978	185.833	265.855	124.529	361.339
(6)	41.759	629.267	133.980	997.125	301.904	2506.928
(7)	19.302	147.079	93.183	687.341	881.058	1168.849
(8)	18.391	171.120	194.827	466.046	2854.875	2756.248
(9)	30.380	850.131	524.838	2046.537	1032.974	19664.687
(10)	247.476	5.099	137.612	176.498	321.142	1661.469
Average Error	60.415	199.437	132.686	498.886	577.265	2934.039

Analysis of Results

➤ We can see that there is a relationship between the number of tweets for a hash-tag and the average error of cross validation. Greater the number of tweets leads to a higher absolute average error for the hash-tag.

➤ In particular it is seen that for each hash-tag the error of one of the cross-validation fold is too high due to the uneven distribution of the data-set. A fold might consider a split wherein the test-data has all high values for the class (tweets during the time of the SuperBowl) and training-data has all low values for the class (tweets before and after the SuperBowl), hence producing a high error value for that fold (e.g. Fold 9 for #gopatriots).

E. Cross Validation with Time Periods

The second part of Question-4 deals with analysis of regression models created for different time-periods during the SuperBowl. Three different time-periods were considered to create the regression models,

1. Before Feb. 1, 8:00 a.m.
2. Between Feb. 1, 8:00 a.m. and 8:00 p.m.
3. After Feb. 1, 8:00 p.m.

Each tweet was segregated based on the time it was posted and split into windows of one-hour. The models were tested using 10-fold Cross Validation and the average errors for all folds obtained were as follows

Table 7.7: Average Error of 10 Fold Cross Validation for each Time-Period

Hash tag	Before	Between	After
#gohawks	167.189	7022.163	2607.692
#gopatriots	16.217	238.102	1760.682
#nfl	75.919	753.944	533.593
#patriots	190.869	93528.077	9745.065
#sb49	39.833	51166.878	12012.449
#superbowl	203.754	12861.877	11834.395

Analysis of Results

- It can be clearly seen that due to the between time-period having only 12 one-hour window the number of instances in this time-period to create a model is very low. Hence the model created is giving very high average error values.
- Since the before time-period has a greater number of instances a better model is created hence giving low average error values.

VIII. CONCLUSION

As proposed an Event-Sequencing and Ad-Celeb Popularity/Comparison Checker was implemented and the results were presented above. The scope of the problem can be further being spread into areas of analytics for advertising agencies and for the celebrity PR teams. Sentiment analysis of the tweets collected can further represent the feelings of an advertisement or a celebrities' performance during the SuperBowl.

REFERENCES

- [1] Bo Wu, Haiying Shen, "Analyzing and predicting news popularity on Twitter", in International Journal of Information Management, 35 (2015) 702–711.
- [2] Shing H. Doong, "Predicting Twitter Hash tags Popularity Level", in 2016 49th Hawaii International Conference on System Sciences, 1530-1605/16 \$31.00 © 2016 IEEE DOI 10.1109/HICSS.2016.247, pp:1959-1968
- [3] Shoubin Kong, Qiaozhu Mei, Ling Feng, Fei Ye, Zhe Zhao, "Predicting Bursts and Popularity of Hashtags in Real-Time", in SIGIR'14, July 06 - 11 2014, Gold Coast, QLD, Australia. Copyright 2014 ACM 978-1-4503-2257-7/14/07 http://dx.doi.org/10.1145/2600428.2609476.
- [4] L.Jaba Sheela," A Review of Sentiment Analysis in Twitter Data Using Hadoop ",in International Journal of Database Theory and Application Vol.9, No.1 (2016), pp.77-86 http://dx.doi.org/10.14257/ijdta.2016.9.1.07
- [5] Anisha P. Rodrigues, Niranjan N. Chiplunkar," Sentiment Analysis of Social Media Data using Hadoop Framework: A Survey", in International Journal of Computer Applications (0975 – 8887) Volume 151 – No.6, October 2016
- [6] Maximilian Jenders, Gjergji Kasneci "**Analyzing and Predicting Viral Tweets**" in WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.ACM 978-1-4503-2038-2/13/05.
- [7] Twitter nlp. 2012.
- [8] Graph databases, http://www.neo4j.org/learn/graphdatabase. 2013.
- [9] neo4j http://www.neo4j.org. 2013.
- [10] Twitter api. 2013.
- [11] Y. Altshuler, W. Pan, and A. S. Pentland. Trends prediction using social di_usion models. In Proceedings of the 5th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction, SBP'12, pages 97{104, Berlin, Hei-delberg, 2012. Springer-Verlag.
- [12] S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R. M. Tripathy, and S. Triukose. Spatio-temporal analysis of topic popularity in twitter. CoRR,abs/1111.2904, 2011.
- [13] L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10, pages 36{44, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [14] L. M. L. Delcambre and G. Giuliano, editors. Proceedings of the 2005 National Conference on Digital Government Research, DG.O 2005, Atlanta, Georgia, USA, May 15-18, 2005, volume 89 of ACM International Conference Proceeding Series. Digital Government Research Center, 2005.
- [15] N. P. Fang Fangl. Detecting Twitter Trends in Real-Time. WITS 2011.
- [16] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, SIGMOD '10, pages 1155{1158, New York, NY, USA, 2010. ACM.
- [17] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, (January):1029–1038, 2010. URL <http://portal.acm.org/citation.cfm?doid=1835804.1835934>.
- [18] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, Tweet, retweet: Conversational Aspects of retweeting on Twitter, volume 0, pages 1–10. IEEE, 2010. URL <http://www.computer.org/portal/web/csdl/doi/10.1109/HICSS.2010.412>.
- [1] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
- [2] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

AUTHORS PROFILE



Krishna Vasu Katarla is a M.Tech student in the Department of Computer Science Engineering at Srivasavi Engineering College, tadepalligudem, India. His research interests include Cloud Computing and Distributed Systems, Data Mining, Big Data Analytics



Dr. Vedula Venkateswara Rao is Professor in the Department of Computer Science Engineering at Srivasavi Engineering College, tadepalligudem, India. He received Masters Degree in Computer Science Engineering from Jawaharlal Nehru Technological University Kakinada, Masters Degree In Information Technology from Punjabi University, Patiala, India and PhD from Gitam University. His research interests include Cloud Computing and Distributed Systems, Data Mining, Big Data Analytics and Image Processing. He published several papers in International conferences and journals.

Invisible Digital Watermarking Methodology for Image Validation

¹ Dr. V.Kakulapati, SNIST, Hyderabad, India, vldms@yahoo.com

²Dr. SR Kattamuri, SNIST, Hyderabad, India, sreeram1203@gmail.com

Abstract— Watermark is the pattern of bits inserted into a digital image that identified the file's copyright information. The word watermarking is originated from faintly visible marks imprinted on organizational stationery. We propose robust and novel strategic in-visible approaches for insertion-extraction of a digital watermark in color images are presented. Contrasting printed watermarks, which are intended to be somewhat visible; the digital watermarks are designed to be completely invisible: the invisible insertion of the watermark is performed in the most significant region of the host image such that tampering of that portion for an intention to remove or destroy will degrade the esthetic quality and value of the image. One feature of the algorithm is that this user defined characters are used as a region of interest for the water marking process and eliminates the changes of watermark removal. Specifically dithering techniques are developed and intended to embed color water marking into the color image. A new technique is proposed and implemented using dithering techniques in various color spaces like RGB, HSV, and CMY. An attempt is made to develop full color water marking the scheme using those techniques.

Key words: - Watermarking; RGB; HSV; CMY; Dither

I. INTRODUCTION

Many researchers has allows digitally representation of watermarking to ascertain itself as a prospective solution for n fortification of rights and regulate information piracy of images . Development of images by using Water marking techniques are classified into two types one is visible and another one is invisible approaches [20]. The visible methods provide indicates for overt assertion of rights with logos and the second methods are presenting covert protection of these rights. Images in first type of methods cannot protected the intellectual property due to easy process of any image in trendy graphic software collections. Whereas, digital water marking come into sight as a tool for secure the multimedia data from copyright violation [8]. Digital water mark will include in turn about the holder and in consequence protect it aligned with illegal use. As the water mark embed with a secret key is utilized to, no one other than the owner will be able to detect this watermark.

Now-a-days. Provision of validity is becoming increasingly significant as more of the world's information is accumulated as enthusiastically moveable bits. This is type of watermarking has been extensively useful to solve copyright issues of digital media concerning to illegitimate utilization or circulation. Numerous gray level image and water marking methods have been suggested [19], but the use of full color watermarks is still not well studied and a major weakness of water marking techniques is that they do not work if the image needs printing .No invisible watermark survives the distortions introduced by printers and any subsequent scanning

does not allow recovery of originally embedded watermark. Hence, the color image watermarking in printed image is focused here. In the present work, we explore the techniques to embed invisible watermark in color images that can survive printed distortions. Printers use a technique called dithering to render images on paper [1]. Dithering involves printing pure color dots on paper in specific patterns determined by the algorithms and dither masks used. Specifically, we investigate dithering techniques to embed color watermark into color image. Here the dithering techniques are proposed and implemented in RGB, HSV, and CMY color spaces.

II. RGB AND HSV COLOR SPACES

In RGB color, space dithering is performed by mapping the colors in the source image colors in RGB set i.e. Red, Green and Blue. A dither mask is issued to map the colors to RGB color sets .The mask is constructing by randomly distributing red, green and blue colors and threshold values are taken by dividing the color range based on the number of tiles for each color inside the mask. This mask is tiled for each color inside the mask .This mask is tiled over the image and dither algorithm is applied .If the image patch value exceeds the threshold value, then the primary color of that particular tile (mask) is retained , otherwise the pixel is made white .The following is the dither mask containing RGB colors .We have chosen the 4X4 mask containing 7 red tiles , 5 green and 4 blue. To avoid patterns in the dithered image, all the three color's tiles are randomly dispersed in the mask. The threshold values in the mask are generated by dividing the color range(0-255) separately for each color channel .Since there are 7 red tiles in the mask, the color range (0 to 255) is divided into 7 part ie. (37,74,110,150,185,220,255) and are randomly placed in the red cells , similarly the color range (0-255) is divided into 5 parts (50,100,150,200,250) for green and 4 parts for blue (64,128,190,255) according to the figure 1.

37	185	80	74
64	150	128	220
250	190	150	100
255	110	200	255

Figure.1: Dither Mask

Algorithm:

1. The compute RGB values of each pixel in the image.
2. Construct a Dither mask (4X4 block) containing randomly distributed RGB values.
3. The Dither mask is tiled over the image and one to one comparison of the corresponding points is done, i.e., pixel values are compared with the threshold values of the mask.
4. If the image pixel value exceeds the threshold value retain the primary color of the otherwise make the pixel white.
5. This process is applied selectively on the channels ie., the threshold values in red tiles are applied on red channels only. Repeat the step 3 till the whole image dithered .The sample images are as shown below in figure 2,3and 4



Figure 2: Image



Figure 3: Dithered image

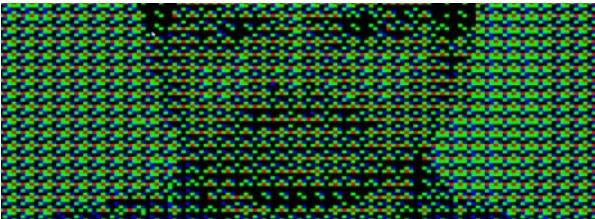


Figure 4: The dithered image zoomed

A. Dithering on the highest Values of RGB

In each pixel of the image, the maximum value of RGB is computed and is compared with threshold value in the mask .If that value is greater than the threshold of the maximum color then the component of that cell is retained. If we consider an example with pixels as Pixel (30, 50,220), Threshold = 150 then the Max (R, G, B) =blue (220) is calculated and Max (RGB)>threshold. Hence blue is retained in the output image as shown in figure 5

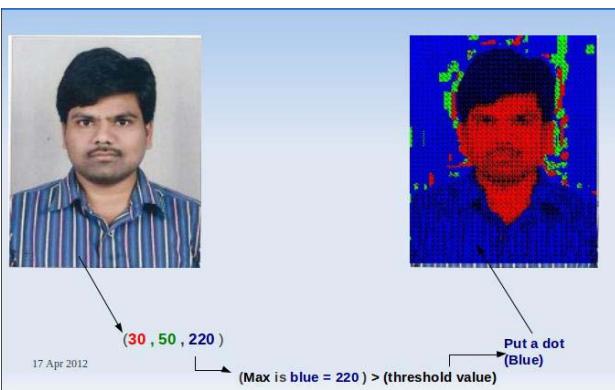


Figure 5 : Dithered to maximum value from primary colors

B .Dithering in HSV space

Hue Dithering

One of the important properties of color is Hue. It refers to dominant wavelength of a specific color. By dithering the image based on 'hue' parameter, the color content of the image can be changed [22].

Algorithm

- 1) Consider an image .Convert each pixel in the RGB image into a HSV.
- 2) Construct a dither mask (4X4) containing randomly distributed hue values.

- 3) Tile, the mask over the image.
 - 4) Perform one of the comparisons of threshold value with the hue value of each pixel of the image.
 - 5) If the hue value of the image exceeds the threshold value, then hue is set as 120 degrees .Otherwise, it is taken as '0'.
 - 6) Repeat step 3 till the whole image dithered.
 - 7) Convert HSV values into RGB primary colors.
- Sample images dithered based on hue are shown below in figure 6,7, and 8



Figure 6: Image



Figure 7: Dithered image



Figure 8: Dithered image containing red (hue=0) and green (hue=120)

In the above figure, hue values are either '0' or '120'. The colors appeared in the dithered image are red and green .The sample image to hue at 120 and 240 are shown below in figure 9, 10 and 11



Figure 9: Image



Figure 10: Dithered image



Figure 11: Zoomed blue (hue=240) and green (hue=120)

C. Dithering on Dispersion

Dithering on Saturation refers to the purity of color. By dithering on saturation the purity of the color can be varied .Hereby considering a dither mask containing random values of saturation 16 parameters, dithering is performed on saturation parameter only. The sample images are as shown below in figure 12 and 13.



Figure 12: Image



Figure 13: Dithered image on 'S'parameters



Image



Dithered image



Color image



Dithered to CMY

D. Dithering on 'Values'

In geometry, the central vertical axis is called value and contains the unbiased, achromatic, or gray colors, value of black and white ranging from 0 and 1. Figure 14 and 15



Figure 14: Image



Figure 15: Dithered image on 'V' Parameter

E. Dithering in CMYK color space

Printing presses and some ink-jet printers utilize four colors (cyan, magenta, yellow and black).Dithering is the most common means of reducing the color range of images .It is one of the principles behind printing technology. The colors are printing by mapping the colors in the source image to CMYK.

C 32	Y 51	M 240	C 200
M 85	C 1	Y 180	M 63
Y 130	Y 196	C 150	Y 164
M 245	C 255	M 94	C 100

Figure 16 : CMYK mask

The color palette contains CMY colors .The figureabove shows the mask that is designed by using 6 cyan tiles, 5 magenta and 5 yellow. The threshold values for cyan are taken by splitting the color range into 6 parts (1,32,100,150,200,255) as shown in figure 16 .Similarly the color range is divided into 5 parts each for yellow and magenta .To map the colors to the CMYK mask , the mask is tiled over the color image and one to one comparison of pixel values of corresponding cells is performed .If the pixel value exceed threshold value, then the color of the corresponding cell is retained .Otherwise white color is retained.

F. Implanting Secret Information

Invisible Digital Image Watermarking

Our algorithm inserts watermark into the color image and utilizes dithering technique to embed watermark in the color image .To make the watermark survive against the distortion created by printers, the image dithered in CMYK color space .To embed secret information two different types of dither masks are using by dithering the image in CMYK color space. For this, a secret code can be embedded in the image.

Embedding Digital watermarking into image

Algorithm

1. For an image converts the primary RGB values of each pixel into subtractive primary CMY colors.
2. Generate two different types of masks (mask-1, mask containing randomly distributed CMY colors).
3. The characters that are to be embedded are converted into ASCII equivalents, the ASCII values are further computed to its bit equivalent.
4. The region where secret information is embedded dithered using two kinds of masks (mask-1 to encode bit 0, mask-2 to encode bit 1.The remaining region of the image dithered using a single mask.
5. If the bit value is '0' mask-1 is tiled over the image .The image is dither with other masks if the bit value is '1'.

Dithering masks

In the embedding the secret code in the image two kinds of masks are used. These masks contain randomly distributing CMY colors and threshold values are generate by dividing the color range separately for each color channel .The following are the dither mask that contains the threshold values of three color's cyan, magenta and yellow. The whole image dithered with a single mask, the place where secret information as to be inserted by dithered using two kinds of masks as shown in the figure17 and 18.

Results :

C 32	Y 51	M 240	C 200
M 85	C 1	Y 160	M 63
Y 130	Y 196	C 150	Y 164
M 245	C 255	M 94	C 100

Figure17: Mask-1

M 20	Y 51	C 252	M 70
C 42	Y 255	M 120	Y 102
Y 153	C 125	M 230	C 168
C 210	M 170	C 83	Y 204

Figure 18: Mask-2

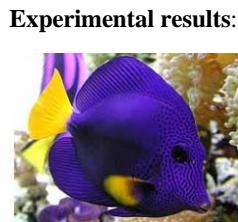


Figure 18: Cover image

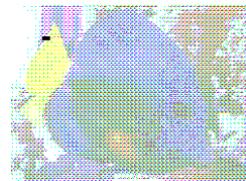


Figure 19: Watermarked image

The black spot in the watermarked image in the above figure 19 is calibrated mark. A secret message is embedded just after the calibrated mark.

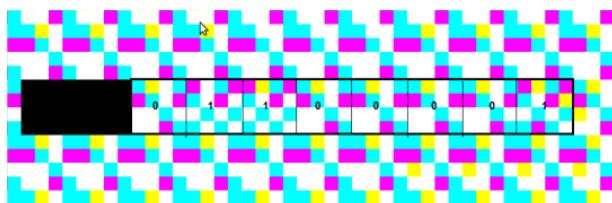


Figure 20: Encoded bits in watermarked image (zoomed)

The figure 20 above shows the embedded bits of character ‘a’. The image is calibrated to find the encoded region. A single character ‘a’ is embedded in the image .The ASCII equivalent of ‘a’ is computed which is ‘97’ and is further converted into bit equivalent i.e. 01100001.These bits encoded in the image using two kinds of masks as shown in figures 20.

Results:



Cover image



Single character encoded image



Image watermarked with 3 characters

Deciphered information from the scanned image

This Algorithm retrieves watermark from the scanned image. The fundamental proposal of dithering is to reduce the color range of the image. The same concept is used in

retrieving secret information .The embedded information can be decoded by dithering the scanned image to CMYK color set.

The Watermarked image Printing and Scanning

The Calibration mark

Initially the watermarked image is calibrated with a black mark near watermarked region, so that embedded information can be easily located. Here we used 4X8 block of black pixels that act as a calibration mark .The watermarked image is printed at different resolutions and scanned at different dpi. The following decoding algorithm is applied on the scanned images to retrieve the secret information.

Algorithm

1. For every pixel in the scanned image compute the mean square distance to each color in CMY color set.
2. Compute the minimum distance of each pixel in the scanned image with the colors in CMYK color space.
3. The scanned image dithered by replacing each pixel in the scanning image to the neighboring color in the CMYK color set.
4. After mapping all colors to CMYK color space, the image is examined near the calibrated mark to find whether the embedded masks are retrieved.

III. EXPERIMENTAL RESULTS



Watermarked image with the calibrated mark

C 32	Y 51	M 240	C 200
M 85	C 1	Y 160	M 63
Y 130	Y 196	C 150	Y 164
M 245	C 255	M 94	C 100

Mask -1

M 20	Y 51	C 252	M 70
C 42	Y 255	M 120	Y 102
Y 153	C 125	M 230	C 168
C 210	M 170	C 83	Y 204

Mask-2

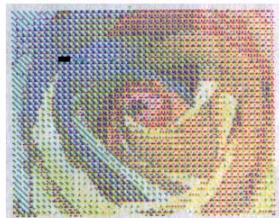
As shown in the figure The above, Mask-1 is used to encode ‘0’ and mask-2 to encode ‘1’.Single character ‘a’ is encoded in the image, its bits equivalent can be observed in the figure above.

Step 1: Encoding the secret characters



Encoded bits (enlarged)

Step 2: Scanning the watermarked image



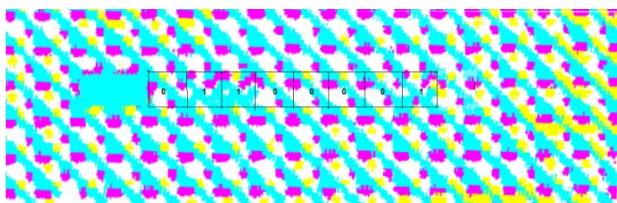
Scanned image at 600dpi

Step 3: The scanned image colors are mapped with CMYK colors.



Scanned the image dithered to CMYK set

Step 4: Examine the region near calibration mark to decode secret bits



Retrieved watermark (decoded bits)

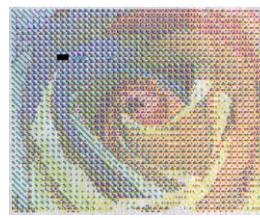
Image at 75 Dpi



Scanned image at 700 dpi

Dithered image.

Image at 1200 Dpi



Scanned image at 1200 dpi



Dithered image

IV. CONCLUSION

In this work, a new method was proposed for embedding secret information invisibly into an image printed on a page. The secret embedding is in the form of dither masks. The pattern of colors printed is unique to the mask and we showed that the pattern is recoverable if scanned at sufficiently high resolution. In our experiment, we got the best results when the scanned resolution is more than four times the print resolution. This technique has the potential to greatly enhance our concept of secure documents.

Bibliography

- [1] V.Ostromoukhov ,P.Emmel , N.Rudaz, I.Amidror, R.D Herch, Multilevel color.halftoning algorithms .Symposium on Advanced Imaging and Network Technologies Berlin ,SPIE Vol 2949 pp. 332-340, oct 1996.
- [2] Nirav Patel ,Binghampton University .Voronoi Diagrams Robust and Efficient implementation Thomas J.Watson School of engineering and Applied Sciences 2005.
- [3] Victor Ostromoukhov Pseudo-Random Halftone screening for color and black and white printing .The 9th International congress in Non-Impact Printing Technologies, Yoohoma ,japan pp .579-582,1993.
- [4] Dr. Chakravarthy Bhagvati,color image processing manual.
- [5] Piyu Tsai,Yu-Chen Hub ; Chin-chen Chang A color image watermarking scheme based on color quantization .A Department of Computer Science and Information Engineering .National Cung Cheng University ,Chiayi ,Taiwan pages 1754-1759,2002
- [6] Gary Bradski , Adrian Kaehler learning open cv manual Computer vision with open cv library
- [7] Joachim J. Eggers, Bernd Girod Quantization effects on Digital Watermarks Telecommunications Laboratory ,University of Erlangen –Nuremberg Stanford University 2000.
- [8] R.G.Van Schyndel A.Z .Tirkel and C.F Osborne , A digital watermark IEEE. Int. Conf.Image Processing vol 2 1994 .pp 86-90.
- [9] W.R. Bender D.Gruhl and N. Morimoto , Techniques for data hiding and Storage and Retrieval of image and video Databases , vol 2420,1995, pp 164-173
- [10] G.K.Wallace, the JPEG still picture compression standard ACM , vol 34 1991 pp 30-44
- [11] F.Hartung B. Girod , Fast public-key watermarking of compressed video in : Proceddings of the IEEE

- International conference on Image Processing Santa Barbara ,CA, USA, October 1997.
- [12] Dariusz Bogumi An assymetric image watermarking scheme resistant against geometrical distortions Institute of Computer , Warsaw University Of technology .Nowowejska 15/19,00-665 Warszawa,Poland June 2005
- [13] M.Kutter,F.Jordan,F.Bossen,Digital signature of color image using amplitude modulation ,in: I. K. Sethi R.Jain(Eds.)Storage and Retrieval for Image and Video Databases V,Vol.3022,SPIE, SanJose,CA,1997,pp.518-526.
- [14] A.Piva,F.Bartolini,V.Cappellini, M.Bamni, Exploiting the cross-correlation of RGB-channels for robust watermarking of color image.processdings of the IEEE International Conference on Image Processing,Vol. 1,
- [15] Campbell,Alastair.The Designer's Lexicon, Chronicle, SanFrancisco,2000
- [16] C.A.Bouman Digital Image Processing January 9,2012.
- [17] Danny Pascale A Review of RGB Color Space from CMY to RGB The Bable Color Company 5700 Hector Desloges Montreal(Quebec) Canada H1T3Z6.2003.
- [18] Scott Cravr, Nasir Memon, Boon-Lock yeo. Resolving Rightful ownership with In-Visible watermarking techniques: limitation, attacks and Implications IEEE may 1998.
- [19] Stephane Derrode, Robust and Efficient Fourier-Mellin Transform Approximations for Gray-Level Image Reconstruction and Complete Invariant Description deBretagne,Technopole,Brest-Iroise.BP832,292858 Brest Cedex,France,31
- [20] Saraju P. Mohanty ,A Dual watermarking technique for images Dept. of Electrical. Engg. Indian institute of science ,Bangalore 1999.
- [21] Victor Ostromukhov. Color in art and science2002.
- [22] Wen chen Yun Qshi,Guorong Xuan, Identifying computer graphics using HSV color model and statistical moments of characteristic functions.
Dept. of Computer Science -China.
- [23] Damin Cardani Adventures in HSV Space dcardani@buena.com.

Designing an Efficient Distributed Algorithm for Big Data Analytics: Issues and Challenges

Mohammed S. Al-kahtani

Dept. of Computer Engg., Prince Sattam bin Abdulaziz University, Saudi Arabia
Email: alkahtani@psau.edu.sa

Abstract - As designing computationally efficient distributed algorithms is very important for analyzing big data this paper presents the current state-of-the-art research in designing distributed algorithms of big data analytics. More specifically, this paper presents a comprehensive survey on existing distributed algorithms of big data analytics – present working principle of existing algorithms, their advantages, limitations and also compare these algorithms in terms of several features. This paper also presents issues and research challenges that may arise to design an efficient distributed algorithm for big data analytics and proposes some solutions to address such challenges. Our research find that these algorithms support parallel processing and designed based on the MapReduce paradigm of big data for a particular application.

Keywords: Big Data, Distributed Algorithm, MapReduce, DBMS, Commodity Hardware.

I. INTRODUCTION

As a large amount of data are being produced from social media, cloud, computer networks, content delivery networks and other emerging technologies and transmitted through Internet Big Data analytics have achieved a widespread popularity. At the same time, analyzing big data has proven to be challenging as traditional computing systems are not able to handle them. Three attributes of big data, velocity, volume and variety, namely 3V also reflects such challenges. Velocity deals with the issue of how quickly large amounts of data are being sent in. Volume deals with the size and amount of data being stored and processed. Variety represents different formats of data, which are mostly unstructured. A good example of 3V attributes of Big Data would be Instagram, which has over 400 million active users with 80 million upload a day on an average. These uploads include multiple formats of pictures and videos [1].

To cope with the challenges big data has emerged, horizontal scaling of a computing system is much more important than the vertical scaling [2]. This means that one computer is not built to be more powerful but the work is spread out over many more less powerful machines. This necessitates the use of distributed algorithms to process big data. Hence, designing efficient distributed algorithms for processing big data is significantly important to achieve computational efficiency.

Several distributed algorithms [1-14] have been designed to process big data. These algorithms are mostly designed to process data of a specific application. These algorithms are also designed based on the MapReduce Paradigm. The MapReduce runs on a cluster of

Lutful Karim

School of ICT, Seneca College of Applied Arts & Technology, Toronto, Canada
Email: lutful.karim@senecacollege.ca

commodity machines and thus, can be used in Hadoop operations and functionalities for large scale data processing. Among many, the Parallel IdeaGraph Algorithm uses MapReduce paradigm to handle big data challenges by implementing a parallel distribution of IdeaGraph [2]. Another Algorithm is Probabilistic Latent Semantic analysis (PLSA) [3] that implements a parallel method to train PLSA under the MapReduce framework. This algorithm addresses the scalability issues in PLSA. Item-based Collaborative-filtering algorithm [4] is a very effective and computationally efficient algorithm in MapReduce Paradigm that uses “hotweight” as the weight [4]. Evolutionary algorithm called the Grouping Genetic Algorithm solves the problem of grouping and is found in the schema optimization in HBase¹. In addition to these algorithms, Locality-aware Scheduling Algorithm (LaSA) performs data locality assignment in Hadoop scheduler to enhance the performance of big data applications [5]. Parallel Two-Pass MDL (PTP-MDL) algorithm, Scalable Nearest Neighbor and Convex Optimization are other algorithms that reduce the computational, storage, and communications bottlenecks [6]. However, these algorithms are application dependent and hence, cannot be used for processing data of all applications. These algorithms only works in commodity machines and cannot be used in low powered wireless nodes such as sensors.

This paper provides a comprehensive survey on distributed algorithms of big data i.e., present working principle of existing algorithms, their advantages, limitations and also compare these algorithms in terms of several features. This paper also presents several challenges and issues to design an efficient distributed algorithm and proposes some solutions for future improvement. The rest of the paper is organized as follow.

Section 2 defines some terminologies and presents MapReduce Paradigm. Section 3 presents a comprehensive review on distributed algorithms of big data. Section 4 analyzes the algorithms, classifies and compares them. Section 5 identifies research challenges to design an efficient distributed algorithm and proposal for improvement. Sections 6 summarizes our research work in this paper.

II. PRELIMINARIES

To understand the working principle of distributed algorithms, the MapReduce framework of Big Data, which has been popularized by Google, is presented first. MapReduce is a scalable and fault-tolerant data processing tool that processes a large amount of data in

parallel with a number of low-end computing nodes [7]. It breaks a large file into a number of small chunks or task and distributes the small unit of works into a number of processing nodes to be executed in parallel. As programs written for MapReduce paradigm are automatically parallelized and can be executed on a large cluster of commodity computers.

The schedule of tasks execution is not predefined in MapReduce. Tasks are scheduled or assigned to nodes during runtime. Moreover, the underlying distributed file system ensures data locality and availability [7]. A few benefits of MapReduce among many are listed below.

- Simple but efficient for query processing in database management system (DBMS).
- It is flexible and easy to use.
- Fault tolerant and highly scalable.
- Can provide it functionalities inside DBMS to support user-defined functions in the relational DBMS.

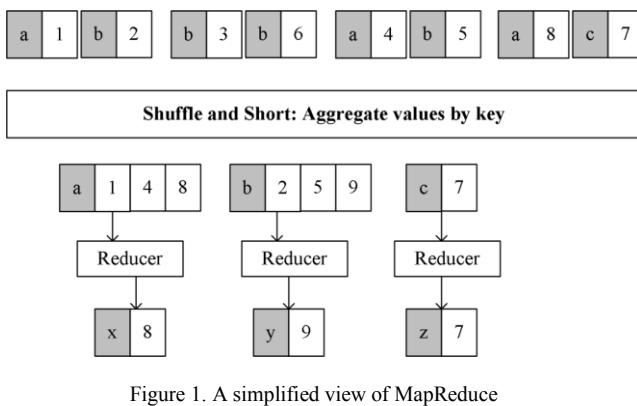


Figure 1. A simplified view of MapReduce

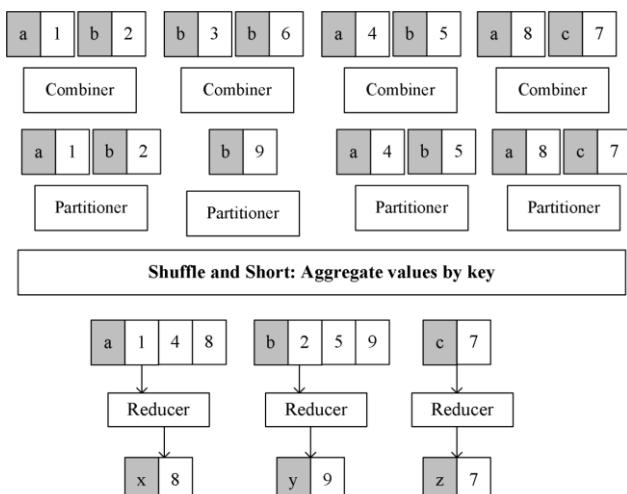


Figure 2. MapReduce with combiners and partitioners

MapReduce uses clustering mechanism on big data analytics. The simplified view of MapReduce is illustrated in Figures 1 and 2. A large file or database gets split into a number of small files when it is stored into the distributed file system of MapReduce [7]. The mapper works onto the every key-value pair and generates intermediate key-value pair. The reducers reduces the number of values by working on the same intermediate

key associated with a number values. Thus, the reducer generates output key value pairs. The general concept is that distributed algorithms based on MapReduce are designed to run on more than one job. For instance, nine jobs run (one at a time) in MapReduce for IdeaGraph algorithm [2]. Figure 3 also illustrates the map and reduces operation process.

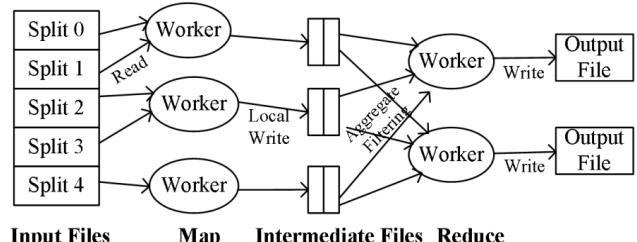


Figure 3. Map and Reduce operation process

III. DISTRIBUTED ALGORITHM IN LITERATURE

Analyzing huge amount of data collected/captured by social media, wireless networks, cloud, and organizations become computationally inefficient unless we design an efficient distributed algorithm to do so. In the following sections, we present several distributed algorithms for big data analysis.

A. IdeaGraph

IdeaGraph is a parallel distributed algorithm that uses MapReduce paradigm and detects chances² where, a chance represents a rare event. The chance discovery (CD) is very significant in human-computer interaction process for quick decision making. IdeaGraph [2] performs well in searching chances. The results generated by IdeaGraph are represented using scenario graph, which can be easily interpreted. IdeaGraph can be both parallel and sequential. The sequential IdeaGraph performs well for small datasets but is not suitable for big data. Though running sequential IdeaGraph on a powerful computer can mitigate the problem to some extent it cannot completely eliminate the problem. This is because a single process has limited computational resources. Thus, the implementation of parallel IdeaGraph that runs into a distributed system such as Apache Hadoop is greatly required.

Thus, the parallel IdeaGraph based on MapReduce is used to improve the performance. The parallel IdeaGraph introduced in [2] is able to discover rare chance from large datasets and also reduces the processing time for big data. The parallel IdeaGraph implements a MapReduce version of IdeaGraph to maintain the accuracy of the original IdeaGraph and achieve better performance. In this implementation of the algorithm, nine jobs run one after another to get the final results. After the initial tests have been run, the authors come to three conclusions. First, the proposed implementation is valid to identify the possibility to get the same results as the sequential IdeaGraph [2]. Second, the parallel implementation consists of several MapReduce jobs and intermediate results are stored into the MapReduce file system (not into the memory). The parallel

implementation does not work well for small data collections as compared to the sequential implementation. Finally, the parallel IdeaGraph implementation is more efficient than the sequential IdeaGraph implementations in terms of processing time and memory cost for large data sets.

B. Probabilistic Latent Semantic Analysis (PLSA)

Probabilistic Latent Semantic Analysis (PLSA) is a powerful statistical technique to analyze relation between co-occurrence data. It has widespread applicability in automated information processing tasks [3]. However, it is often difficult and time consuming to train large datasets. This scalability problem of PLSA can be addressed by implementing a parallel method to train PLSA under the MapReduce computing framework. P2LSA and P2LSA plus are two types of parallel P2LSA that are designed based on MapReduce paradigm in [3]. The P2LSA executes E-step and M-step similar to the Map and Reduce functions, respectively. A large amount of data is transferred between E and M-steps to train PLSA models and perform the operations (i.e., read and process the input records). This overloads the network and increases the overall processing time. The P2LSA plus performs two different jobs to finish the whole task, but that reduces the degree of parallelism. The work done in [3] introduces an algorithm that works differently from these two MapReduce algorithms. In this approach, the authors implement the EM algorithm and then parallelize it. Thus, this algorithm is more straightforward and applicable to many areas.

C. Recommendation Algorithm

The idea of recommendation system was first introduced in 1994 by a research team who launched a system called GroupLens and set up a collaborative filtering engine [4]. Recently, the recommendation systems that use recommendation algorithms have achieved widespread applicability in commercial video, audio streaming and e-commerce applications such as YouTube, Netflix video recommendation system, Amazon, ebay. Hence, the recommendation systems also require processing huge amount of data and information as a part of big data analytics. Thus, the traditional operating mode of recommendation system does not work well because of the limited resources to process big data [4]. Hence, Hadoop is being used as the key to solve this problem as the distributed infrastructure of Hadoop can fully utilize the storage and computing ability. This leads to another algorithm, which is a collaborative filtering recommendation algorithm running on a Hadoop platform using “hotweight” [12] as the weight. The “hotweight” is the average rating of each item taken from historical data of users. This greatly improves the efficiency of a system when dealing with huge amounts of data. Traditional algorithms cannot recommend an item if a user is added to the system (as item vectors are all zeros). However, collaborative algorithms that use hotweight can recommend items for new users, which are basically the bestselling products. The computation of hotweight uses MapReduce computation model, i.e., distributed framework available in Hadoop. The process

starts by collecting key-value pairs with the same key. By using this distribute system, the computational efficiency is significantly improved [4]. The hotweight calculation process is shown in Figure 4.

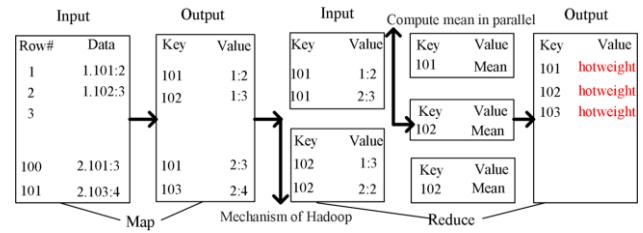


Figure 4. MapReduce process of computing hotweight

D. Locality-aware Scheduling Algorithm (LaSA)

Many data intensive computing applications such as MapReduce, HBase, and Hadoop [5] have been built for big data processing. Data access, placement and computations pose a great challenge in big data framework. These problems can be alleviated by increasing data locality in distributed environment [5]. MapReduce uses Hadoop as a distributed file system, process scheduler, execution environment and framework. However, the process scheduler fails to account for data locality issues. Thus, the work done in [5] introduces Locality-aware Scheduling Algorithm (LaSA) algorithm that enhances data locality assignment in Hadoop scheduler and also increases the performance of data-intensive computing applications. The end point of LaSA is to accomplish locality-aware resource assignment in order to decrease the bottleneck of transmission by following the weight of data interference. The concept of data interference weight is made known to MapReduce framework by the LaSA as well as the locality-aware scheduler is introduced in JobTracker. The data interference weight of each node that depends on resource assignment is calculated by the LaSA. Then, the job tracker selects a node with the least weight in order to execute the task, which in turn reduces the congestion of network transmissions.

E. Parallel Two-Pass Minimum Description Length (PTP-MDL)

In data storage and communications, distributed cloud computing and big data frameworks introduce new challenges. For instance, data streaming uses distributed computing. Data is processed remotely in clusters and results are streamed through a network to the user. The use of data streaming in big data framework makes it imperative to use fast lossless data compression algorithms whose primary features include good compression quality and high throughput [8]. The PTP-MDL algorithm is used for such data compression. It possesses the random access property. Hence, if a part of a file is compressed using PTP-MDL that part can be decompressed without decompressing the whole file. Also the PTP-MDL algorithm has numerical results that provide a better trade-off between compression and throughput. Thus, this algorithm is very useful and efficient for big data problems.

The Parallel Two-Pass MDL (PTP-MDL) can compress certain data files or part of files while decompresses

others. The block diagram of the PTP-MDL encoder is displayed in Figure 5 as well the block diagram of the PTP-MDL decoder in Figure 6.

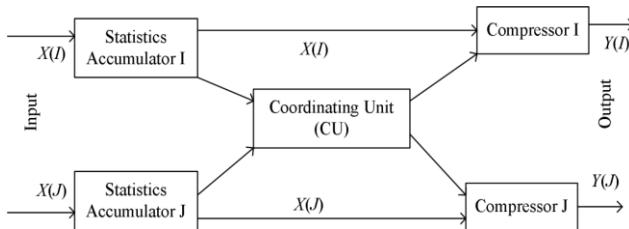


Figure 5. PTP-MDL encoder

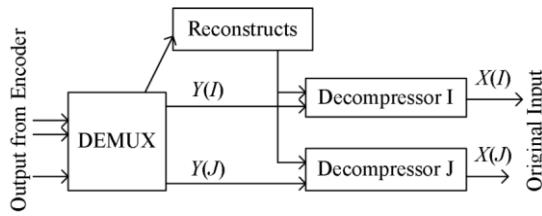


Figure 6. PTP-MDL decoder

F. Scalable Nearest Neighbor Algorithm

Many computer vision algorithms are computationally inefficient as they search for the most similar matches to high-dimensional vectors. This is also known as nearest neighbor matching. Scalable nearest neighbor algorithms [9] are efficient as they can find the nearest neighbors in large data sets very fast. However, the use of large training sets is significantly important to achieve better performance in these algorithms [9]. Though many applications improve processing time they are subject to settle for an approximate search i.e., after performing the search the neighbors returned as search result are not all exact. Although it is an approximation they are still very close to the exact neighbors. It provides almost 100% correct neighbors. Moreover, it is much faster than linear search (two or more orders of magnitude faster).

The Scalable Nearest Neighbor algorithms are able to match the nearest neighbor very fast in a large data set and thus, it improves the overall performance of network bandwidth. The priority search k -means tree algorithm is one of the very fast nearest neighbor algorithms. The search tree is built using k -mean clustering. This clustering algorithm partitions the data points at each level into k individual clusters/regions and recursively applies the same approach to the points in each region. The recursion terminates when the number of points in a region becomes smaller than k .

G. Convex Optimization Algorithm

The main focus of convex optimization algorithms for big data [6, 11] is to reduce the computational, storage, and communication bottlenecks. Convexity in signal processing dates back to the dawn of the field, with problems like least squares (LS) being ubiquitous across nearly all subareas [6]. The importance of convex formulations and optimizations has increased even more dramatically in the last decade because of the rise of new

theory for structured sparsity and rank minimization, and successful statistical learning models such as support vector machines [6]. Due to the popularity of convex optimization, the algorithms are considered greatly to accommodate large data sets and also to solve problems in unprecedented dimensions.

H. Other Algorithms

The Item-Based collaborative filtering recommendation algorithm (IBCF) computes the product of co-occurrence matrix and user item vector (which also estimates the complexity of this algorithm). Traditional algorithms require huge storage for large matrix. The method in [4] is designed to save space to store sparse matrix that is used by IBCF. This method is just storing the nonzero data with the coordinates. In matrix multiplication, the output matrix $C[m \times k]$ is computed by two input matrices $A[m \times n]$ and $B[n \times k]$. The computation of each element in matrix C is independent of each other. Hence, in Map phase, all the elements from matrix A and matrix B should be collected together under the same key to calculate the element of matrix C . Then the Reduce function calculates every element of matrix C in parallel [4]. The matrix multiplication process is shown below in Figure 7.

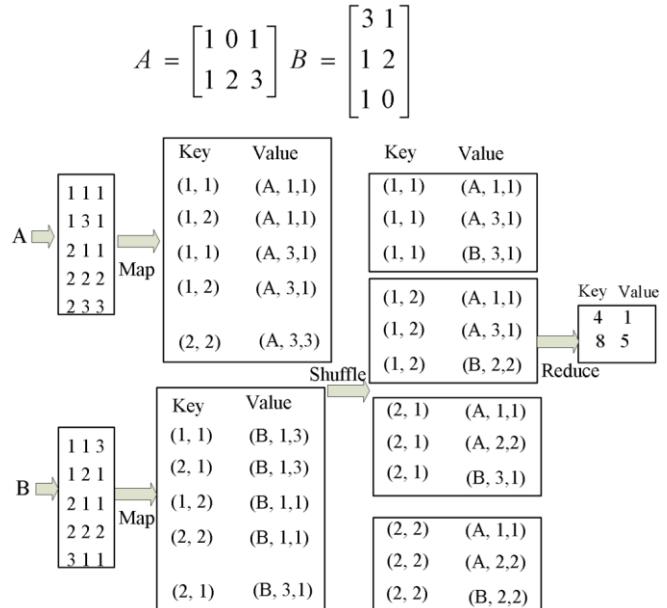


Figure 7. MapReduce process of matrix multiplication

Operations of an evolutionary algorithm are very important, which conducts the direction of evolution to optimized solutions. An algorithm with efficient operations can let solutions jump from a local area to search in the global space, which leads the evolution to a better and faster convergence point [1]. For this HBase optimizing problem, genetic operations together with validity check for evolution process use mutation and crossover. A mutation operator adds or removes one or more columns inside one CF. Redundant columns can be removed to make a CF more effective for query types that do not need those columns. Moreover, some other

Table 1. Comparison of existing distributed algorithms

Features	Idea-Graph	PLSA	Recommen-dation Algorithm	LaSA	PTP-MDL	Nearest Neigbo r	Convex Optimization
Parallel processing	√	√	√	√	√	√	√
Sequential processing	√	X	X	X	X	X	√
Uses MapReduce paradigm	√	√	√	√	√	√	X
Statistical Model	X	√	X	X	X	X	√
Use high throughput lossless data compression	X	X	X	X	√	X	X
Use collaborative filtering engine	X	X	√	X	X	X	X
Used in computer vision algorithms for searching the most similar matches to high-dimensional vectors	X	X	X	X	X	√	X
Support locality-aware resource assignment	X	X	X	√	X	X	X
Used in structured sparsity and rank minimization	X	X	X	X	X	X	√
Able to match the nearest neighbor very fast in a large data set	X	X	X	X	X	√	X

columns can be added so that the CF can match more query types [1]. The crossover operator exchanges columns between two CFs inside an HBase schema. Such crossover mixes good sub-solutions of CFs without any disruption of the partitions. The split up operator is designed to split a CF into two smaller CFs inside a solution. A diving point is chosen by the operator, and the columns before diving point are kept in the current CF. The columns after the dividing point form a new CF. Among other algorithms, the work done in [13] introduces distributed structured prediction learning algorithm that reduces processing time and storage requirements and the work done in [14] presents a comprehensive survey on in-memory big data management and processing such as fault tolerance and consistency of in memory environment.

IV. PERFORMANCE ANALYSIS

In section 3, we presented a number of distributed algorithms for processing big data. Most of these algorithms are based on MapReduce paradigm as it provides scalability and fault tolerance for big data. However, MapReduce cannot substitute DBMS rather it can complement DBMS with scalable and flexible parallel processing³. However, input and output costs of MapReduce still require to be addressed for successful implications.

Parallel IdeaGraph based on MapReduce framework can handle big data challenges. It has been found that parallel IdeaGraph is better than sequential counterpart for processing huge volume of data. On the other hand, sequential IdeaGraph is much better for processing small data set. In general, the parallel IdeaGraph implementation performs better than the sequential IdeaGraph implementations in terms of data processing time and cost of memory [2].

The Probabilistic Latent Semantic Analysis (PLSA) parallel algorithm is another popular distributed algorithm of big data. The performance of the PLSA is evaluated in terms of speedup which is defined as the ratio of standalone running time to the Hadoop cluster running time. Three benchmarked datasets: 20Newsgroups dataset, its subset mini-Newsgroups, and a crawled

document dataset [3] were used to evaluate its performance. The speedup is high when the dataset is small. However, the speed up does not increase linearly for small dataset with the increasing number of processors because of the communication costs between the Map phase and the Reduce Phase. In addition, the initialization and the partition processes also take some time. The speedup increases almost linearly as the number of the processors for large datasets. Results on performance evaluation of PLSA [3] conclude that the PLSA training method on MapReduce can deal with large datasets efficiently and hence provides a practical solution to big data analysis applications.

The LaSA is a job scheduler algorithm for Hadoop-MapReduce. It improves the data locality problem of Job Tracker in the original MapReduce framework [5]. The LaSA arranges the resource assignment to avoid required data missing by using the weight of data interference concept. On the other hand, the compression ratio of the PTP-MDL algorithm for real-world data is comparable to existing universal data compressors. Moreover, the throughput of PTP-MDL scales well with the number of parallel units.

The scalable nearest neighbor algorithms can search for the nearest neighbor in a high dimensional vector space very fast. Thought this makes the algorithms computationally inefficient they are widely used in many computer vision and machine learning algorithms [8]. As distributing the search into multiple machines requires less memory access/overhead their performance gets better.

An Evolutionary Algorithm was also presented that finds the optimum column family schema for a given set of user queries. The reading performance of the optimized column family schema was evaluated using a real dataset, which contains 2.6 million rows of aggregated tracking data and 1.3 million user queries [1]. It is found that the reading performance of HBase is improved using optimized column family schema. User queries from a testing set shows that the average response time is reduced by up to 72% compared to un-optimized column family schemas [1]. Table 1 compares several distributed

algorithms that we presented in this paper in terms of a number of features.

V. RESEARCH CHALLENGE AND PROPOSED SOLUTIONS

After analyzing existing distributed algorithms of big data that are presented in previous sections, the following challenges/limitations have been identified.

- They are mostly designed based on MapReduce paradigm for a specific application or context and thus, cannot efficiently be used for other applications.
- They are not dynamic. They are mostly designed for large amount of data. Hence, it is not efficient for small amount of data.
- They are designed considering that processing is distributed into a large number of commodity computers. However, they do not consider distributed processing at data collecting devices such as sensors.
- They support either sequential or parallel implementations. They do not support both.
- Statistical analysis is a great tool for big data analysis and hence, can be considered an integral part of an efficient distributed algorithm of big data analytics. Most of them do not support statistical models.
- They cannot extract rare but important events (for making decisions) from huge data unless they are designed to do so.
- Moreover, partitioning highly correlated data to maximize data locality and reduce communication cost is a great challenge.

These research challenges can be alleviated by introducing a distributed algorithm for big data framework that is application independent and can be applied to any big data application. The proposed algorithm is scalable, dynamic and fault tolerant, i.e., it supports (i) as many nodes as needed, (ii) both parallel and sequential implementation based on the amount of data to process (i.e., if the amount of data exceed a certain threshold it works in parallel mode; otherwise, it should work in sequential mode to reduce processing overhead) (iii) continuous operation even if any node fails. Moreover, the algorithm should work in different types of nodes ranging from low powered RFID tags and sensors to high speed/powered computers. The proposed algorithm also integrates a statistical model to validate the results.

VI. CONCLUSION

We presented a literature survey on the existing distributed algorithms for big data analytics. We compared these algorithms based on different features and identified their limitations. We also identified research challenges to design an efficient distributed algorithm for big data analytics and proposed some solutions. This research work will lead us to design an efficient distributed algorithm for big data analysis in terms of processing time, memory requirements in future.

REFERENCES

- [1] Yang F, Milosevic D, Cao J. An Evolutionary Algorithm for Column Family Schema Optimization in HBase. *IEEE First International Conference on Big Data Computing Service and Applications (BigDataService)*, Redwood City, CA, 2015, pp. 439-445.
- [2] Wang Q., Wang H, Zhang C, Wang W, Chen and F. Xu. A Parallel Implementation of Idea Graph to Extract Rare Chances from Big Data. *2014 IEEE International Conference on Data Mining Workshop*, Shenzhen, 2014, pp. 503-510.
- [3] Liang Z, Li W and Li Y. A Parallel Probabilistic Latent Semantic Analysis method on MapReduce Platform. *2013 IEEE International Conference on Information and Automation (ICIA)*, Yinchuan, 2013, pp. 1017-1022.
- [4] Lu F, Hong L, Changfeng L. The improvement and implementation of distributed item-based collaborative filtering algorithm on Hadoop. *2015 34th Chinese Control Conference (CCC)*, Hangzhou, 2015, pp. 9078-9083.
- [5] Chen TY, Wei HW, Wei MF, Chen YJ, Hsu TS and Shih WK. LaSA: A locality-aware scheduling algorithm for Hadoop-MapReduce resource assignment. *2013 International Conf. on Collaboration Technologies and Systems (CTS)*, San Diego, CA, 2013, pp. 342-346.
- [6] Cevher V, Becker S, Schmidt M. Convex Optimization for Big Data: Scalable, randomized, and parallel algorithms for big data analytics. In *IEEE Signal Processing Magazine*, Sept. 2014, 31(5), pp. 32-43.
- [7] Maitrey S, Jha CK. Handling Big Data Efficiently by Using Map Reduce Technique. *2015 IEEE International Conference on Computational Intelligence & Communication Technology (CICT)*, Ghaziabad, 2015, pp. 703-708.
- [8] Krishnan N, Baron D. A Universal Parallel Two-Pass MDL Context Tree Compression Algorithm. In *IEEE Journal of Selected Topics in Signal Processing*. June 2015, 9(4), pp. 741-748.
- [9] Muja M, Lowe DG. Scalable Nearest Neighbor Algorithms for High Dimensional Data. in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Nov 2014, 36(11).
- [10] Facchinei F, Scutari G, Sagratella S. Parallel Selective Algorithms for Nonconvex Big Data Optimization. In *IEEE Transactions on Signal Processing*, April 2015, 63(7), pp. 1874-1889.
- [11] Daneshmand A, Facchinei F, Kungurtsev V, Scutari G. Hybrid Random/Deterministic Parallel Algorithms for Convex and Nonconvex Big Data Optimization. In *IEEE Transactions on Signal Processing*, August 2015, 63(15), pp. 3914-3929.
- [12] Ghuli P, Ghosh A, Shettar R. A collaborative filtering recommendation engine in a distributed environment. *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, Mysore, 2014, pp. 568-574.
- [13] Schwing AG, Hazan T, Pollefeys M, Urtasun R. Distributed Structured Prediction for Big Data. *NIPS Workshop on Big Learning*, Nevada, USA, Dec 2012.
- [14] Zhang H, Chen G, Ooi BC, Tan KL, Zhang M. In-Memory Big Data Management and Processing: A Survey. In *IEEE Transactions on Knowledge and Data Engineering*, 2015, 27(7), pp. 1920-1948.

A Survey on Memory Virtualization in Cloud

PVSS Gangadhar
Scientist-D& Research
Scholar,
NIC, Meity, Govt of India,
Dept. of IT, GIT,
Gitam University
Visakhapatnam,
Andhra Pradesh, India
pvss.gangadhar@nic.in

Dr.Ashok Kumar Hota,
Scientist-F,
NIC, MEITY,
Odisha state Unit,
Bhubaneswar
Govt of India
ak.hota@nic.in

Dr.M.Venkateswara Rao
Professor, Dept of IT, GIT,
Gitam University,
Visakhapatnam,
Andhra Pradesh, India
mandapti_venkat@yahoo.com

Dr.V.Venkateswara Rao
Professor, Dept. of CSE,
Sri Vasavi Engg College
Tadepalligudem,
Andhra Pradesh, India
venkatvedula2017@gmail.com

Abstract—the Cloud Computing Infrastructure-as-a-Service (IaaS) layer offers a service for on demand virtual machine images (VMIs) operation. This service offers a elastic platform for cloud users to build up, install, and test their applications. Virtual machine environments are becoming more universal due to the augmented performance of service hardware and the materialization of cloud computing for huge scale applications. Because of the reason that virtual machines continues to grow, performance critical applications will need proficient methods to realize their tasks. The operation of a VMI classically engages booting the image, installing and configuring the software packages. In the conventional approach, when a cloud user requests a new platform, the cloud provider chooses a suitable template image for cloning and deploying on the cloud nodes. The template image encloses pre-installed software packages. If it does not fit the requirements, then it will be tailored or the new one will be formed from scratch to fit the request. In the context of cloud service management, the customary approach features the difficult issues of handling the complexity of interdependency between software packages, scaling and maintaining the deployed image at runtime. The cloud providers would like to automate this process to improve the performance of the VMIs provisioning process, and to give the cloud users more flexibility for selecting or creating the appropriate images while maximizing the benefits for provider's intern of time, resources and operational cost. The increasing demand for storage and computation has driven the growth of large data centers—the massive server farms that run many of today's Internet and business applications.

A data center can comprise many thousands of servers and can use as much energy as a small city. The massive amounts of computation power contained in these systems results in many interesting distributed systems and resource management problems. In this paper we focus to investigate challenges related to data centers, with a particular emphasis on how new virtualization technologies can be used to simplify deployment, improve resource efficiency, and reduce the cost of reliability, all in application agnostic ways. We first study problems that relate to the initial capacity planning required when deploying applications into a virtualized data center, issues related to memory utilization among virtual machines and performance metrics for memory virtualization..

Keywords-Cloud Computing; IaaS; Virtual Machine Image(VMI); Data Center; Memory Virtualization; Capacity planning;

I. INTRODUCTION

All Cloud Computing is a model for enabling service users to have ubiquitous, convenient and on- demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services), that can be rapidly provisioned and released with minimal management effort or service-provider interaction. The Cloud will provide IT similar to public utilities providing electricity, gas, and water. There is no need to have to own the hardware & the staff. There will be multiple public cloud providers. The generalized architecture of cloud computing is given in figure-1 below.

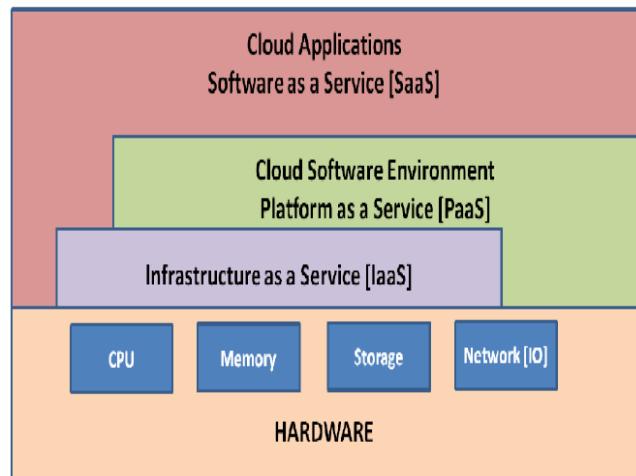


Figure1: Cloud computing Architecture

Infrastructure as a service refers to the sharing of hardware resources for executing services, typically using virtualization technology. With this so called Infrastructure as a Service (IaaS) approach, potentially multiple users use existing resources. The resources can easily be scaled up when demand increases, and are and are typically charged for on a per -pay-use basis. In the Platform as a Service (PaaS).approach, the offering also includes a software execution environment, such as an application server. In the Software as a Service approach (SaaS), complete applications are hosted on the Internet. Virtualization is the process of decoupling the hardware from the operating system on a physical machine. Virtualization can be thought of essentially as a computer within a computer, implemented in software. This is true all the way down to the

emulation of certain types of devices, such as sound cards, CPUs, memory, and physical storage. The Figure2 below describes how the physical resources like servers, storage and network are separated from the operating systems and the applications using the virtual infrastructure.

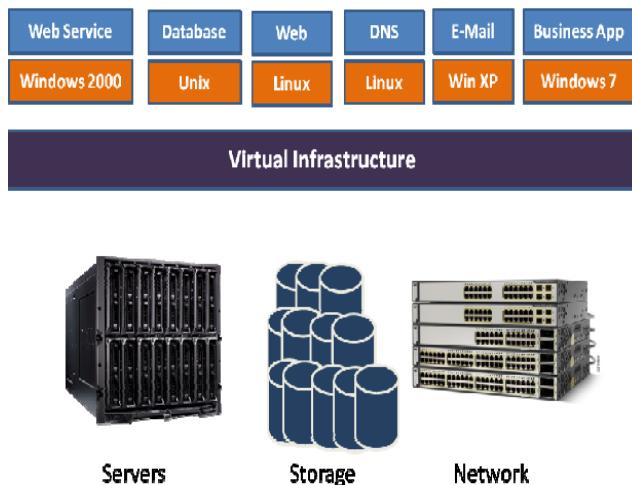


Figure2: Hardware / Software Separation in a Virtualized Environment

Cloud Computing takes Virtualization to the Next Step. Virtual Machines & services can be rented as needed from a cloud service provider. However for the management of such a process requires that the resources be managed in an efficient way. Resource Management is this process of managing the Physical Resources like CPU, Memory, Network etc across various Virtual Machines (VM) based on policies. The figure-3 below gives a frame work of the cloud infrastructure with resource orchestration.

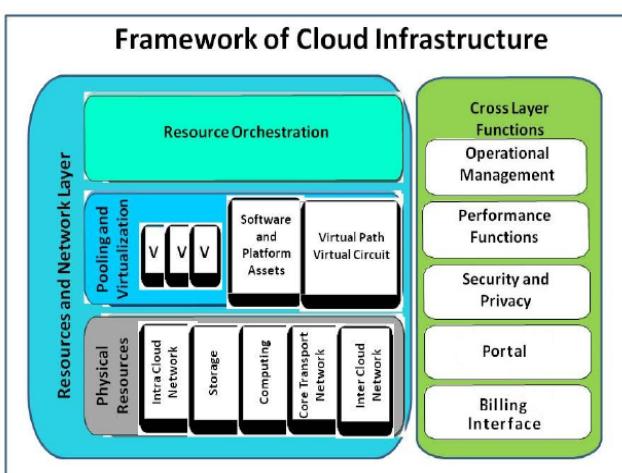


Figure3: Framework of cloud infrastructure with resource orchestration.

II. VIRTUALIZATION

Virtualization is the process of decoupling the hardware from the operating system on a physical machine. An instance of an operating system running in a virtualized environment is known as a virtual machine. Figure-4 below gives architecture of the virtualization with multiple virtual machines sharing the same hardware resources

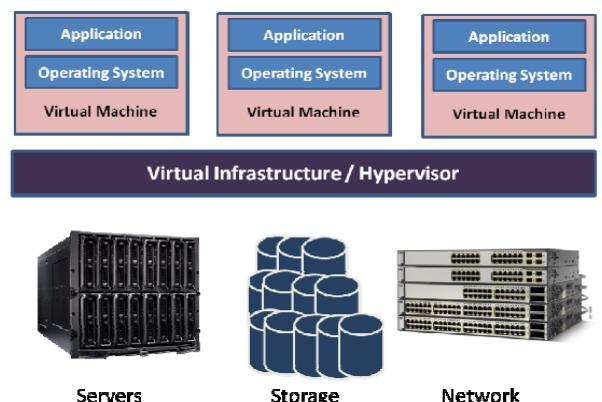


Figure 4 Virtualization Architecture

Virtualization technologies allow multiple virtual machines, with heterogeneous operating systems to run side by side and in isolation on the same physical machine. By emulating a complete hardware system, from processor to network card, each virtual machine can share a common set of hardware unaware that this hardware may also be being used by another virtual machine at the same time. The operating system running in the virtual machine sees a consistent, normalized set of hardware regardless of the actual physical hardware components.

Virtualization is adopted in cloud architecture because of the following reasons

- Hardware independence: The guest VM Sees the same Hardware regardless of the host Hardware
- Isolation – VM's operating system is isolated from the host operating system
- Encapsulation–Entire VM encapsulated into a single file
- Simplified administration because of Hardware
- Increased hardware utilization, Server consolidation, Decreased provisioning, independence & portability times and Improved security
- The other reasons include reduced capital expenditure, reduced operating expenditure, reduced risks of data outage and reduced energy consumption

A. Virtualization terminology:

Host Machine: A host machine is the physical machine running the virtualization software. It contains the physical resources, such as memory, hard disk space, and CPU, and other resources, such as network access, that the virtual machines utilize.

Virtual Machine: The virtual machine is the virtualized representation of a physical machine that is run and maintained by the virtualization software. Each virtual machine, implemented as a single file or a small collection of files in a single folder on the host system, behaves as if it is running on an individual, physical, non-virtualized PC.

Virtualization Software: Virtualization software is a generic term denoting software that allows a user to run virtual machines on a host machine.

B. Components of virtualization:

Guest OS: A guest OS is an operating system that runs in a virtual environment. A guest OS may be a client desktop, physical server or any other operating system that runs independently of dedicated hardware resources. Instead, the guest OS uses hardware resources allocated dynamically through a hypervisor or similar intermediary software.

Hypervisor or virtual machine manager (VMM): A hypervisor also called a virtual machine manager (VMM), which is a program that allows multiple operating systems to share a single hardware host. Each operating system appears to have the host's processor, memory, and other resources all to itself. The task of this hypervisor is to handle resource and memory allocation for the virtual machines, ensuring they cannot disrupt each other, in addition to providing interfaces for higher level administration and monitoring tools [8].

A virtual machine monitor monitors a system of virtual machines (sometimes called hardware virtual machines), which allow the sharing of the underlying physical machine resources between different virtual machines, each running its own operating system. A virtual machine monitor monitors the software layer providing the virtualization, which is called a virtual machine. The VMM is the control system at the core of virtualization. It acts as the control and translation system between the VMs and the hardware.

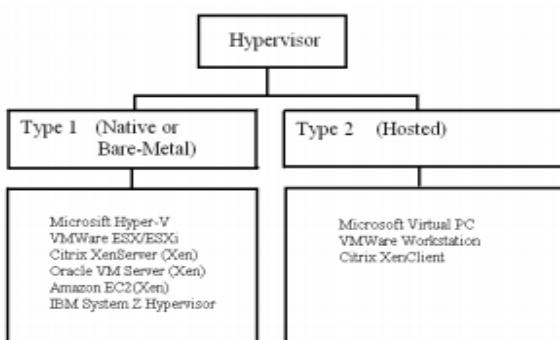


Figure 5 hypervisor types

III. RESOURCE MANAGEMENT

Resource Management is the process of managing the Physical Resources like CPU, Memory, Network etc across various Virtual Machines (VM) based on policies. The figure-6 below gives positioning of resource management in the management and distributed services in virtualization architecture.

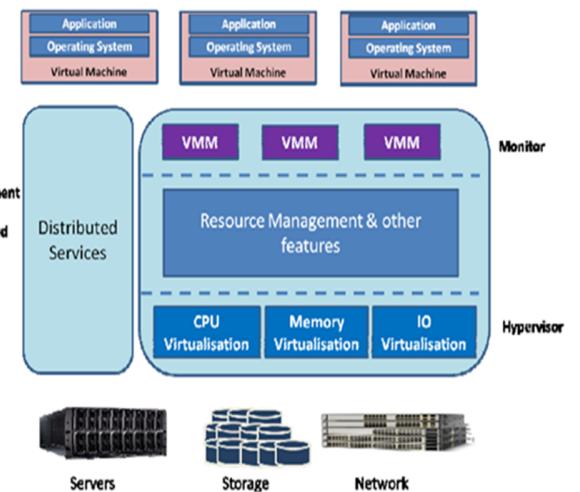


Figure 6: Management & Distributed Services in Virtualization Architecture

The Goal of resource management is three fold.

- Performance isolation: This prevent VMs from monopolizing resources and Guarantees predictable service rates
- Efficient utilization: This is achieved through exploiting under committed resources and over commit with graceful degradation
- Support flexible policies: Absolute service-level agreements are met and relative importance of VMs is controlled efficiently.

A. Capacity monitoring

The key metrics monitored for capacity planning are:

- (i) Server utilization: Peak/average server resource utilization – memory /CPU/resource, server bottlenecks and correlation with a number of users/VMs.
- (ii) Memory usage: Memory utilization on each server, capacity bottlenecks and relationship with number of users/VMs and with different cloud services.
- (iii) Network usage: Peak/average network utilization, capacity/bandwidth bottlenecks and relationship with a number of users/VMs and with different cloud services.

(iv) Storage utilization: Overall storage capacity metrics, VM/virtual disk utilization, I/O performance metrics, snapshot monitoring and correlation with a number of users/VMs and with different cloud services.

IV. MEMORY VIRTUALIZATION

It divides the physical memory, allocate memory for virtual machine instances when starting up, and release memory from virtual machines when shutting down. Every running instance of OS sees a continuous memory space and is isolated from the memory space of other instances. The hypervisors are capable of memory address conversion from the guest instance physical memory address to the machine physical address. The operating system of a running instance maps the application virtual memory to guest instance physical memory. It can increase the memory allocation to a guest OS at a later stage. The virtualization software has commit feature for the virtual machines.

The operating system maps the virtual page numbers to physical page numbers that are stored in page tables. All modern x86 CPUs contain a memory management unit (MMU) and a translation look aside buffer (TLB) for improvement virtual memory. In order to run multiple virtual machines on a system, another level of memory virtualization is essential. Thus, one has to virtualized the MMU to support the guest OS. The guest OS controls the mapping of virtual addresses to the guest memory physical addresses, but the guest OS cannot have access to the actual machine memory. The TLB hardware is used by the VMM to map virtual memory directly to the machine memory in order to avoid the two levels of translation on every access.

A. Memory Management

Memory manager (MM) such as Hypervisor or ESXi in VMware uses five memory management mechanisms. Page sharing, ballooning, memory compression, swap to host cache, and regular swapping to dynamically reduce the amount of machine PM required for each VM. Virtualization causes an increase in the amount of PM required due to the extra memory needed by MM for its own code and for data structures.

This additional memory requirement can be separated into two components

- A system-wide memory space overhead for the VM kernel and various hosts.
- An additional memory space overhead for each VM.

The amounts of memory reserved for these purposes depend on a variety of factors, including the number of virtual CPUs, the configured memory for the guest operating system,

whether the guest operating system is 32-bit or 64-bit, and which features are enabled for the virtual machine.

B. Re-Claiming unused Memory

The memory manager of the hypervisor detects whether the virtual memory is actually used by the guest OS or not. If not, the hypervisor shall be able to assign the unused part of the memory to another guest OS, so that the memory can be shared among the guest OS. Hence this feature is required for memory over-commitment.

This is achieved through traditional method of adding transparent swap layer or using an implicit co-operation. Ballooning is a method where Guest OS manages memory implicit cooperation by paging in / out of the virtual disk. In Page Sharing, multiple VMs running same OS de-duplicate redundant copies of code, data etc.

C. Non Uniform Memory Access scheduling

Periodic rebalancing of the memory usage computes VM entitlements & memory locality, assign “home” node for each VM and migrate VMs and pages across nodes VM migration is to move all VCPUs and threads associated with VM and migrate to balance load and improve locality Page migration allocates new pages from home node and carries migration and replication.

D. Implementations of Memory Virtualization

Application-level integration – Applications executing on connected computers openly connect to the memory pool through an Application Programming Interface(API)

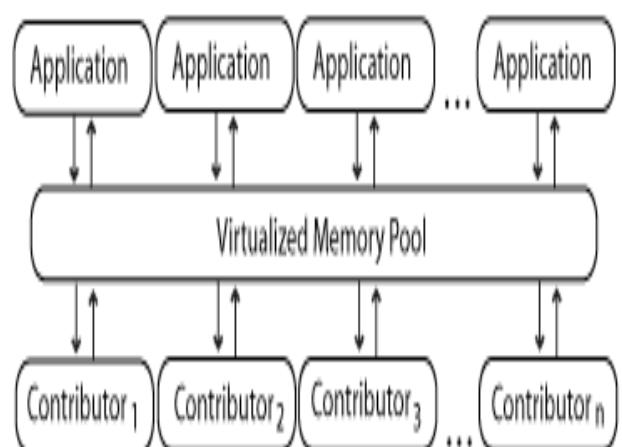


Figure 7: Memory Virtualization through Application Level Integration

Operating System Level Integration – The operating system first connects to the memory pool, and makes that pooled memory available to applications.

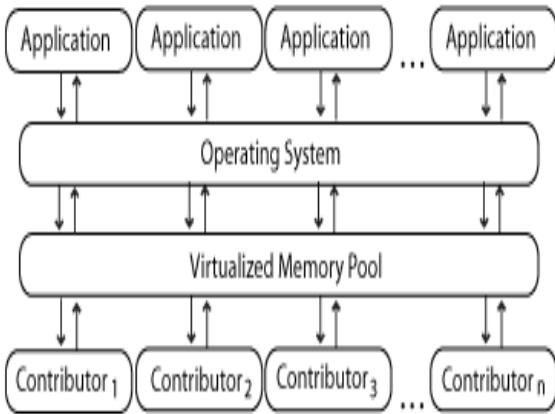


Figure 8: Memory Virtualization through Operating System Level Integration

V. CONCLUSION

Data center and desktop computing successfully use virtualization for better utilization of computing capacity, to balance computing load, manage complexity and parallelism and improve security by isolation. However Virtualization may not work well for Resource-intensive applications where VMs may have RAM/CPU/SMP limitations or situations where custom hardware devices are required. Some hardware architectures or features are impossible to be virtualized as certain registers or states are not exposed. Mobile and embedded computing currently lag behind virtualization since most hypervisors only support the x86 platform, require large memories, have poor real-time support and are inefficient with microkernel Operating System's.

REFERENCES

- Q. Ali, "Scaling web 2.0 applications using Docker containers on vSphere 6.0," <http://blogs.vmware.com/performance/2015/04/scaling-web-2-0-applications-using-docker-containers-vsphere-6-0.html>, 2015, (Accessed on 03/21/2016).
- J. Almeida, V. Almeida, D. Ardagna, C. Francalanci, and M. Trubian, "Resource Management in the Autonomic Service-Oriented Architecture," in Proceedings of the 2006 IEEE International Conference on Autonomic Computing (ICAC 2006), June 2006, pp. 84–92.
- J. Anselmi, E. Amaldi, and P. Cremonesi, "Service Consolidation with End-to-End Response Time Constraints," in Proceedings of 34th Euromicro Conference on Software Engineering and Advanced Applications (SEAA 2008.), September 2008, pp. 345–352.
- M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica et al., "A view of cloud computing," Communications of the ACM, vol. 53, no. 4, pp. 50–58, 2010.
- B. Arnold, S. A. Baset, P. Dettori, M. Kalantar, I. I. Mohomed, S. J. Nadgowda, M. Sabath, S. R. Seelam, M. Steinder, M. Spreitzer, and A. S. Youssef, "Building the ibm containers cloud service," IBM Journal of Research and Development, vol. 60, no. 2-3, pp. 9:1–9:12, March 2016.
- M. Assuncao, M. Netto, B. Peterson, L. Renganarayana, J. Rofrano, C. Ward, and C. Young, "CloudAffinity: A framework for matching servers to cloudmates," in Proceedings of the 2012 IEEE Network Operations and Management Symposium (NOMS 2012), April 2012, pp. 213–220.
- J. Banks, J. S. Carson, B. L. Nelson, and D. M. Nicol, Discrete-event system simulation, Prentice Hall, 2010.
- P. Barham et al., "Xen and the art of virtualization," in Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP 2003), October 2003, pp. 164–177.
- L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," Computer, vol. 40, no. 12, pp. 33–37, Dec. 2007.
- C. L. Belady and D. Beaty, "Roadmap for datacom cooling," ASHRAE journal, vol. 47, no. 12, p. 52, 2005.
- A. Beloglazov and R. Buyya, "Energy efficient allocation of virtual machines in cloud data centers," in Proceedings of the 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGrid 2010), May 2010, pp. 577–578.
- "Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers," in Proceedings of the 8th International Workshop on Middleware for Grids, Clouds and e-Science, December 2010, pp. 4:1–4:6.
- F. Caglar, S. Shekhar, and A. Gokhale, "A performance Interference-aware virtual machine placement strategy for supporting soft realtime applications in the cloud," Institute for Software Integrated Systems, Vanderbilt University, Nashville, TN, USA, Tech. Rep. ISIS-13-105, 2013.
- R. Calheiros and R. Buyya, "Energy-efficient scheduling of urgent Bag-of-Tasks applications in clouds through DVFS," in Proceedings of the 6th IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2014), December 2014, pp. 342–349.
- R. N. Calheiros, M. A. Netto, C. A. De Rose, and R. Buyya, "EMUSIM: an integrated emulation and simulation environment for modeling, evaluation, and validation of performance of cloud computing applications," Software: Practice and Experience, vol. 43, no. 5, pp. 595–612, 2013.
- M. Chen, H. Zhang, Y.-Y. Su, X. Wang, G. Jiang, and K. Yoshihira, "Effective VM sizing in virtualized data centers," in Proceedings of the 2011 IFIP/IEEE International Symposium on Integrated Network Management (IM 2011), May 2011, pp. 594–601.
- Y. Chen, S. Alspaugh, D. Borthakur, and R. Katz, "Energy efficiency for large-scale MapReduce workloads with significant interactive analysis," in Proceedings of the 7th ACM European Conference on Computer Systems, April 2012, pp. 43–56.
- Y. Chen, A. S. Ganapathi, R. Griffith, and R. H. Katz, "Analysis and lessons from a publicly available Google cluster trace," University of California, Berkeley, Tech. Rep., 2010.
- P. DELFORGE, "Energy efficiency, data centers—NRDC," <http://www.nrdc.org/> energy/data-center-efficiency-assessment.asp, (Accessed on 02/18/2016).
- B. H. Li, X. Chai, and L. Zhang, "New advances of the research on cloud simulation," in *Advanced Methods, Techniques, and Applications in Modeling and Simulation*, vol. 4 of *Proceedings in Information and Communications Technology*, pp. 144–163, 2012.
- S. Jafer, Q. Liu, and G. Wainer, "Synchronization methods in parallel and distributed discrete-event simulation," *Simulation Modelling Practice and Theory*, vol. 30, pp. 54–73, 2013.
- R. Fujimoto, A. Malik, and A. Park, "Parallel and distributed simulation in the cloud," *SCS Modeling and Simulation Magazine*, pp. 1–10, 2010.
- A. W. Malik, A. J. Park, and R. M. Fujimoto, "An optimistic parallel simulation protocol for cloud computing environments," *SCS M&S Magazine*, vol. 4, 2010.
- A. Javor and A. Fur, "Simulation on the Web with distributed models and intelligent agents," *Simulation*, vol. 88, no. 9, pp. 1080–1092, 2012.
- IEEE Std 1516.1-2010, *IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)*, Framework and Rules Specification, 2010.

26. IEEE Std 1516.2-2010, *IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)*, Object Model Template (OMT) Specification, 2010.
27. IEEE Standard, *1516.1-2010—IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)—Federate Interface Specification*, 2010.
28. Google, "Google App Engine", (2012), [online]. Available: cloud.google.com [Nov 1, 2012].
29. Amazon, "Amazon Elastic Compute Cloud (Amazon EC2)", (2012), [online]. Available: aws.amazon.com/ec2/ [Nov 1, 2012].
30. Microsoft, "Windows Azure.", (2012), [online]. Available: windowsazure.com [Nov 1, 2012].
31. IBM, "SmartCloud." (2012), [online]. Available: ibm.com/cloudcomputing [Nov 1, 2012].
32. P. Mell and T. Grance, "The NIST definition of cloud computing(draft)," *NIST special publication*, vol. 800, p. 145.
33. A. Desai, "Virtual Machine." (2012), [online]. Available: <http://searchservervirtualization.techtarget.com/definition/virtualmachine>
34. VMWare, "vSphere ESX and ESXi Info Center.", (2012), [online]. Available: vmware.com/products/vsphere/esxi-and-esx [Nov 1, 2012].
35. Microsoft, "Windows Virtual PC.", (2012), [online]. Available: <http://www.microsoft.com/windows/virtual-pc/> [Nov 1, 2012].
36. Xen, "Xen Hypervisor.", (2012), [online]. Available: <http://www.xen.org/products/xenhyp.html> [Nov 1, 2012].
37. Microsoft, "Hyper-V Server 2012.", (2012), [online]. Available: microsoft.com/server-cloud/hyper-v-server/ [Nov 1, 2012].
38. KVM, "Kernel-based Virtual Machine.", (2012), [online]. Available: linux-kvm.org [Nov 1, 2012].
39. Oracle, "VirtualBox.", (2012), [online]. Available: virtualbox.org [Nov 1, 2012]



Dr. Vedula Venkateswara Rao is Professor in the Department of Computer Science Engineering at Srivasavi Engineering College, tadeppalligudem, India. He received Masters Degree in Computer Science Engineering from JawaharLalNehru Technological University Kakinada, Masters Degree in Information Technology from Punjabi University, Patiala, India and PhD from Gitam University. His research interests include Cloud Computing and Distributed Systems, Data Mining, Big Data Analytics and Image Processing. He published several papers in International conferences and journals.

Author Biographies



P.V.S.S.GANGADHAR is Scientist-D in NIC & Ph.D Scholar. Presently studying Ph.D, Department of Information Technology at Gitam Institute of Technology, Gitam University, Vishakapatnam, and AndhraPradesh, India. His Research interests e-governance, Cloud computing, fuzzy logic and Data mining.



Dr. Ashok Kumar Hota is Scientist-F at NIC, MEITY,OSU, Bhubaneswar, Govt of India. His research interests include e-governance, Tribal informatics, Cloud Computing, Data Mining, and Big Data Analytics. He published several papers in International conferences and journals



Dr. Mandapati Venkateswara Rao is Professor in Department of Information Technology at Gitam Institute of Technology, Gitam University, Vishakapatnam, India. He has received M.Tech in CST and PhD in Robotics from Andhra University. His Research Interests includes Robotics, Cloud Computing and Image processing. He published several papers in International conferences and journals.

A Comparative Study on MAC Protocols for Wireless Sensor Networks on Energy Reduction

Rajan Sharma, Research Scholar,
Department of Electronics and Communication
Engineering
K. Gujral Punjab Technical University, Jalandhar
Punjab, India
E-mail: rajansharma.ece@gmail.com

Balwinder Singh Sohi
Department of Electronics and Communication
Engineering
CGC Group of Colleges
Punjab, India
E-mail: bssohi@yahoo.com

Abstract—Wireless Sensor Networks (WSNs) leads to a most versatile solution in numerous applications such as smart building, tracking the targets and many more. Basically, WSNs comprises of high count of inexpensive sensor nodes those are scattered in target area for gathering the desired information. Medium Access Control (MAC) Protocols plays key role for energy efficient operation in WSNs as it manages radio communication on the shared medium. WSN is multidisciplinary area of research and has increased acceptance because of its application. Due to lower sensing range and changing topology there is needed to achieve an efficient medium access for energy efficiency. Variety of MAC protocols have been proposed for WSNs, based upon different objectives like energy efficiency, delay, throughput and packet loss. In recent years, WSNs has gained special attention in research community. This paper outlines the properties of WSNs that plays an important role for designing of MAC layer protocol highlighting pros and cons. Finally, we investigate on various MAC protocols designs for WSN.

Keywords-Wireless Sensor Networks; MAC; Energy Reduction; Energy Efficiency

I. INTRODUCTION

Wireless Sensor Networks (WSNs) is an amalgamation of large number of sensing nodes with one or more base stations. In sensor network, sensing nodes are deployed randomly to monitor the physical parameters such as temperature, sound, pressure, etc. Sensor nodes are autonomous petite devices that comprises of components such as sensing module, processing module, radio unit and power unit as presented in Figure 1. Sensor nodes capture, gather analyze data and communicate with other sensor nodes through radios. The gateways exchange gathered data with applications [1]. These nodes basically sense, compute and communicate the data to other nodes or base station. It preserves the association to the outer world where the activities such as data management, data evaluation, and data grant provide to the effective approaches.

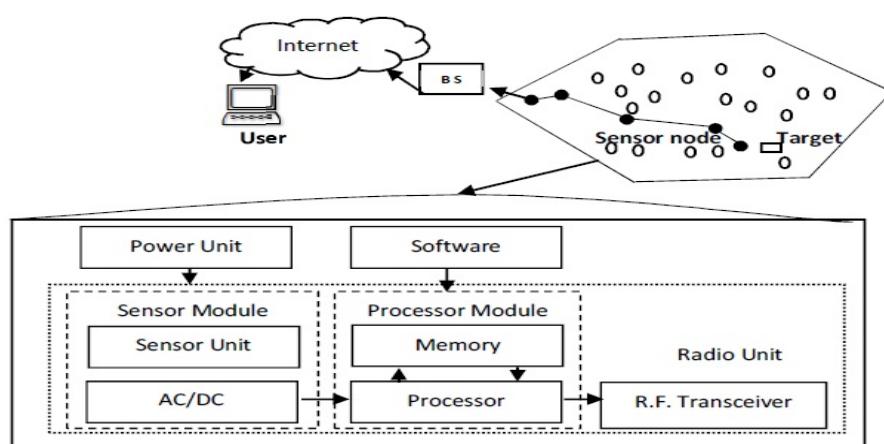


FIGURE 1: BASIC STRUCTURE OF WIRELESS SENSOR NETWORK

Wireless Sensor Network (WSN) provides applications for various fields like defense, environment monitoring, disaster relief, target tracking and detection etc.

The data link layer operates between physical layer and network layer. The data link layer is further segmented into Logic Link Control (LLC) sub layer and Medium Access Control (MAC) sub layer. The Control Mechanism for channel access at MAC layer influence significantly on energy efficiency, performance and lifetime of the network as it is responsible for fair and efficient access of shared medium by nodes, resolving conflicts between competitor nodes and correcting errors. MAC sub layer provides various other functions like dentation, addressing, frame recognition and delimiting, gathering information of source station and transferring the data [2]. According to 802.3 functions MAC needs to fulfill are: padding, checking errors in the frame sequence, transferring and receiving of frames, appending and checking frame check sequence and adding and removing preamble [3].

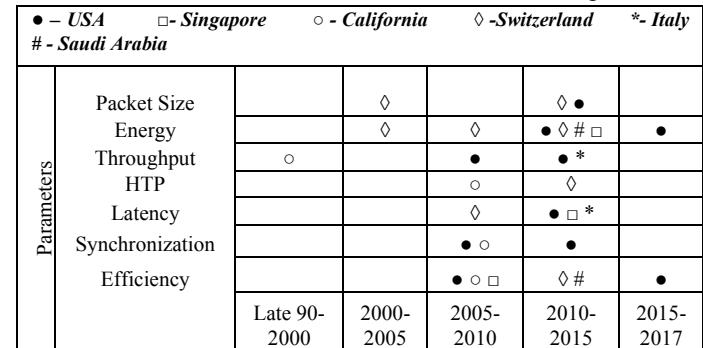
WSNs Medium Access Control (MAC) protocol design requirements are entirely different from the conventional wireless networks. The prime design requirement in WSNs is to achieve higher energy efficiency but in the case of conventional wireless computer networks, higher throughput is needed. Since only source of energy for sensors nodes are batteries which are inconvenient to replace so saving of energy is primary goal for design of MAC layer protocols. As sensor nodes are battery powered, these nodes face energy depletion issues because of which packets are delayed or abort. MAC protocols are designed with an aim to

reduce power consumption and to prolong network lifetime. Hence, we need the operations to be energy efficient so that the sensors lifetime can be extended. Out of sources of power consumption (sensing, radio operation and computation of data) the radio operation consumes maximum power of sensor nodes [4]. Reducing energy consumption with a aim to achieve higher energy efficiency are the significant aspect that can increase the battery life of mobile terminals which is limited [5]. There are many protocols which have been designed for wireless sensor networks. The protocols which are related to the wireless sensors should have an awareness related to the energy to increase the life of the network [6]. The sources of energy wastage in WSNs that occurs due to the following reasons:

- Overhearing:* Sometimes nodes gather the packets that were supposed to be intended to other nodes on the shared medium which wastes the energy; later leads to energy depletion [7]. These sensor nodes may run out of power while they need to be in active state.
- Packet Overheads:* The size of control packets (headers and trailers) associated with data should be short and exchanging handshaking packets should be limited as they consume lot of energy.

- Collision:* Collision of packets either due to retransmissions or discarding of packet requires more energy of the nodes in comparison to the normal transmission.
- Idle Listening:* Sensor nodes sometimes waste energy when they are active in accessing channel but they have no packets to listen.

The major goal in designing MAC protocol is to minimize these issues to save power and to achieve higher energy efficiency [8]. Section II discuss the Prior work for various types of MAC protocol used in WSN that provide solution for power consumption reduction with other constraints of sensor nodes. Section III shows the Valuation of MAC protocols



discovered, followed by providing the comparisons and conclusion in Section IV.

FIGURE 2: MAJOR CONTRIBUTION BY VARIOUS COUNTRIES IN THE FIELD

II. PRIOR WORK

The work in the respective field has always been on its track. Thus, the analysis of the major contributed countries in the golden years of the field are discussed in Figure 2.

Kulkarni [9] explains, the major challenges faced by the sensor nodes, those are the problem of early sleeping, synchronization and overheads of packets. In some of the protocols energy consumption was improved yet need to be compromise on other metrics like packet delay and throughput of the network. IEEE 802.15.4 provides preamble based protocol which can save the energy but face the problem related to acknowledgement of preamble.

W. Ye et al. [10] provides sensor MAC (S-MAC), protocol that manages periodic sleep and listen synchronization. The message passing concept where packet is divided into frames before sending reduces communication overheads, which further saves the energy. The sleep schedule helps in saving energy of idle listening. The variable traffic load creates problem in Listen and sleep period which leads to a decrease in efficiency of algorithm. Overhearing issues are still there because of the collision probability which was also faced in Halkes [17].

El-Hoiydi [11] proposed Spatial TDMA and CSMA with preamble sampling protocol where the nodes use different channel for data and control. TDMA method is used for data channel whereas CSMA method is used for control channel. The preamble along with data is sent with an aim to reduce the power consumption by finding active and in-active state of nodes but the size of preamble may be the disadvantage.

A.El-Hoiydi et al. [12] gave Wise MAC, protocol that is quite similar to CSMA but use a single channel to reduce the idle listening. Dynamic preamble is used to reduce the power consumption over fixed length preamble. This helps to manage the variable traffic conditions. Hidden terminal Problem creates issue in efficiency of the algorithm. This algorithm suffers from higher latency and increase in power consumption because of buffering of broadcasting packets for finding active and in active sensor nodes.

Rajendran et al. [13], proposed a TDMA based TRAMA (Traffic Adaptive MAC), that build the vitality effectiveness by wiping out concealed terminal issue and guarantees that the hubs display at one jump from the transmitter will get the bundles without the impacts. Time is categorized into schedule access and random-access periods. Less collision probability and more sleep time helps in achieving better results but the timing slots creates issue of synchronization.

Bao et al. [14] proposed, node activation multiple access (NAMA) where time slots are divided and nodes present at two hop from the transmitter is elected to avoid the collision. This technique faces problem of delay due to more sleep time. Tay Y.C et al. [15], proposed event driven sensing environment. SIFT is MAC based protocol that uses probability distribution function that is non-uniform. The latency and energy consumption was taken into consideration but idle listening was still there because of slot listening while sending even the overhearing problem was forced. So, this technique increases complexity.

Alazzawi and Elkateeb [16], focuses on features of WSN are dense deployment multi-hop routing, self-organizing and dynamic topology. The algorithm divides time into slots and assigns those to the sensing nodes depending upon data gathering tree. Though the latency is reduced but collision could be avoided.

Halkes et al. [17] proposed T-MAC to handle variable traffic load. Time out MAC (T-MAC) provided better result as the nodes sleep when periods end listens, but synchronization can create problem of early sleeping.

Lin et al. [18] proposed Dynamic Sensor MAC (DS-MAC), that include feature of dynamic duty cycle that helps in reducing latency. In synchronization periods, the sensing nodes have one hop latency value. When this value is high it reduces twice so that neighboring nodes are not affected which hence saves the power consumption.

Safwat et al. [19], provides integrated MAC layer protocol with routing protocol. Cui et al. [20] integrates MAC protocol with routing protocol at physical layer where the TDMA scheme of variable length helps in energy utilization but the node distance creates the complexity. Ding et al. [21] proposed MINA that integrates routing protocol with MAC in multi hop infrastructure network architecture; the sensing nodes are grouped under base stations which are at one hop distance. This head of cluster will decide the schedule for data transmission, but this technique is in tolerant to the failure.

Zorzi et al. [22] proposed Geographic Random Forwarding (GeRaF) which combines routing protocol with MAC protocol using CSMA/CA technique. Rugin et. al. [23] improved by reducing up to one channel system, but still the latency issue was there due to transmission request during awake_time.

Roy and Sharma [24] proposed AEEMAC that is Adaptive energy efficient MAC protocol that uses duty cycle to avoid idle listening and reusing the channel by adaptive sleeping. This protocol by protocol with MAC protocol using CSMA/CA technique.

Protocol	Technique	Communication pattern support	Adaptively to change	Time needed for synchronization
S-MAC	CSMA	All	Good	No
T-MAC	CSMA	All	Good	No
DS-MAC	CSMA	All	Good	No
Wise-MAC	Np-CSMA	All	Good	No
D-MAC	TDMA/ Slotted ALOHA	Coverage based	Good	No
TRAMA	TDMA	All	Good	Yes
SIFT	CSMA/ CA	All	Good	No

TABLE 1 ASSESSMENT OF MAC PROTOCOLS

This protocol shows better results than ESMAC that was proposed by S. Hayat [25], but still complexities in data and acknowledgment packets where there. Kwan-WuChin et al. [26] provides E2MAC that focus on providing better results for packet delivery rate and delay in Packets.

Bhuiyan et. al [27] proposed Intelligent MAC (I-MAC) that works on the basis of intelligent sleep and wakeup schedule and is inspired from 802.11 DCF. Cano et al. [28], proposed MAC protocol in which preamble of IEEE 802.15.4 is used to find the sleep time for the sensor node. This process creates packet overhead reducing the efficiency of the protocol.

N. Javaid [29] reduced the ideal listening time with proposed protocol by extending MAC header. The sensor nodes periodically check the channel so that they could increase their sleep time which provides energy efficiency.

Later Escolar et al.[30] discussed that due to dynamic change of topology the energy saving is the major issue in WSN. A lot of research has been done to provide protocols for this

FIGURE 3: PARAMETRIC CONTRIBUTION BY VARIOUS

Parameters	Packet size	[28]		[19][12]	[28]
Energy Consumption		[10]	[15]	[9][10][12][18]	[19]
Throughput			[21]	[9][16][22][23]	
HTP	[24]	[27]	[13]	[12]	[29]
Latency		[26]	[15]	[12][18][10]	[30]
Synchronization	[25]		[13][17]		
Efficiency		[10]	[13][17]	[10]	[19]

RESEARCHERS IN THE MAJOR YEARS OF RESEARCH

purpose. Most of the protocols aim to achieve energy reduction by focusing on data transmission. This provides reduction in packet loss as well as packet delay. Figure 3 depicts the complete parametric contribution by various researchers in the golden era of the field.

III. THE ASSESSMENT OF MAC PROTOCOLS EXPLORED

Table 2. Provides comparison of various MAC protocol based upon parameters like communication pattern, time needed for synchronization and adaptability to the changes. The Table 2 and Figure 4 provides the comparison between ESAC, E2MAC, IMAC and AEEMAC protocols on the basis of data and acknowledgement packets which are sent during the time slots.

Protocol	Data packets	Acknowledgement packets	Total	Total/min
ES-MAC	490	490	980	16
E2MAC	240	240	480	8
AEEMAC	100	100	800	16
I-MAC	80	80	170	2

TABLE2. COMPARISON OF DIFFERENT MAC PROTOCOL

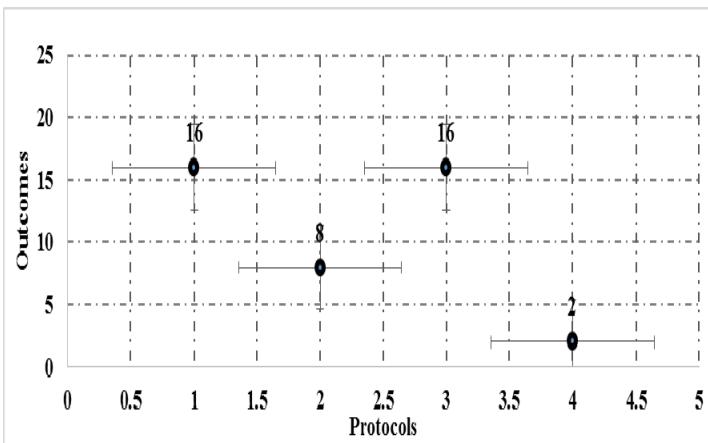


FIGURE 4. COMPARISON OF MAC PROTOCOLS BASED UPON NUMBER OF PACKETS

IV. CONCLUSION

Wireless Sensor Networks (WSNs) have become popular due to its exceptional capabilities & numerous applications in many fields such as military, industries, health sector and surveillance etc. Various MAC protocols available in the literature have been reviewed in this paper. Power saving of WSNs is considered as the prime aspect in designing MAC protocols because it is not feasible to change the batteries of sensor nodes in many applications wherever there is less human intervention. The responsibility of the MAC protocol is to manage radio communication fairly and efficiently on shared medium. In this survey, we found that the main target is to keep the sensor nodes in power saving mode to the maximum extent and to reduce the parameters which are source of energy inefficiency such as re-transmission due to collision or congestion, idle channel sensing, overhearing, and overhead due to control messages so that lifetime of network can be increased. This review will be beneficial as it highlights to the possibilities of the development of some new energy efficient MAC approaches to extend lifetime of the WSNs.

ABBREVIATIONS

WSN	Wireless Sensor Network
MAC	Medium Access Control
LLC	Logic Link Control
HTTP	Hyper Text Transfer Protocol
SMAC	Sensor Medium Access Control
TDMA	Time Division Multiple Access
CSMA	Carrier Sense Multiple Access
TRAMA	Traffic Adaptive Medium Access
NAMA	Node Activation Multiple Access
T-MAC	Time Out Medium Access Control
AEEMAC	Adaptive Energy Efficient Energy Protocol
I-MAC	Intelligent Medium Access Control
E2MAC	Energy Efficient Medium Access Control
ES-MAC	Energy Efficient Sensor Medium Access Control

REFERENCES

- [1] S. Kurt, H. U. Yildiz, M. Yigit, B. Tavli, and V. C. Gungor, "Packet Size Optimization in Wireless Sensor Networks for Smart Grid Applications," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2392–2401, 2017.
- [2] N. Zaman, L. T. Jung, and M. M. Yasin, "Enhancing Energy Efficiency of Wireless Sensor Network through the Design of Energy Efficient Routing Protocol," vol. 2016, 2016.
- [3] S. Zhuo, Z. Wang, Y. Q. Song, Z. Wang, and L. Almeida, "A Traffic Adaptive Multi-Channel MAC Protocol with Dynamic Slot Allocation for WSNs," *IEEE Trans. Mob. Comput.*, vol. 15, no. 7, pp. 1600–1613, 2016.
- [4] K. Pattamasiriwat and C. Jaikaeo, "Evaluation of Low Power Listening MAC Protocol on Network Monitoring in Wireless Sensor Networks," pp. 0–5, 2017.
- [5] M. Dong, K. Ota, and A. Liu, "RMER: Reliable and Energy-Efficient Data Collection for Large-Scale Wireless Sensor Networks," *IEEE Internet Things J.*, vol. 3, no. 4, pp. 511–519, 2016.
- [6] M. Dong *et al.*, "Evaluation of Low Power Listening MAC Protocol on Network Monitoring in Wireless Sensor Networks," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 0–5, 2016.
- [7] J.-D. Decotignie, C. Enz, V. Peiris, and M. Hübner, "WiseNET: An Ultra Low-Power Concept for Wireless Sensor Networks," *Tech. Mess. - Platf. für Methoden, Syst. und Anwendungen der Messtechnik*, vol. 77, no. 2, pp. 107–112, 2010.
- [8] V. K. Sachan, "Energy Efficient Wireless Sensor Networks using Co-operative MIMO : A Technical Review," vol. 135, no. 11, pp. 20–27, 2016.
- [9] S. S. Kulkarni, "TDMA service for sensor networks," in *24th International Conference on Distributed Computing Systems Workshops, 2004. Proceedings.*, 2004, pp. 604–609.
- [10] W. Ye, J. Heidemann, and D. Estrin, "Sleeping for Wireless Sensor Networks," vol. 12, no. 3, pp. 493–506, 2004.
- [11] A. El-Hoiydi, "Spatial TDMA and CSMA with preamble sampling for low power ad hoc wireless sensor networks," in *Proceedings - IEEE Symposium on Computers and Communications*, 2002, pp. 685–692.
- [12] A. El-Hoiydi and J. D. Decotignie, "WiseMAC: An ultra low power MAC protocol for multi-hop wireless sensor networks," *Algorithmic Asp. Wirel. Sens. Networks*, no. 5005, pp. 18–31, 2004.
- [13] V. Rajendran, "Energy-efficient medium access control for wireless sensor networks," *Wirel. Networks*, pp. 445–452, 2006.
- [14] L. Bao and J. J. Garcia-Luna-Aceves, "A New Approach to Channel Access Scheduling for Ad Hoc Networks," *Comput. Eng.*, pp. 210–220, 2001.
- [15] Y. C. Tay, K. Jamieson, and H. Balakrishnan, "Collision-minimizing CSMA and its applications to wireless sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 6, pp. 1048–1057, 2004.
- [16] L. K. Alazzawi, A. M. a. M. Elkateeb, and A. Ramesh, "Scalability Analysis for wireless sensor networks Routing Protocols," *22nd Int. Conf. Adv. Inf. Netw. Appl.*, pp. 139–144, 2008.
- [17] G. P. Halkes, T. V. A. N. Dam, and K. G. Langendoen, "Comparing Energy-Saving MAC Protocols for Wireless Sensor Networks," pp. 783–784, 2005.
- [18] P. Lin, C. Qiao, and X. Wang, "Medium Access Control With A Dynamic Duty Cycle For Sensor Networks," in *Proc. {IEEE} Wireless Commun. and Networking Conf. (WCNC)*, 2004, vol. 3, pp. 1534–1539.
- [19] A. Safwat, H. Hassanein, and H. Mouftah, "ECPS and E2LA : New Paradigms for Energy Efficiency in Wireless Ad hoc and Sensor Networks," *GLOBECOM '03. IEEE Glob. Telecommun. Conf. (IEEE Cat. No.03CH37489)*, pp. 3547–3552, 2003.
- [20] S. Cui, R. Madan, A. J. Goldsmith, and S. Lall, "Joint Routing, {MAC}, and Link-Layer Optimization in Sensor Networks with Energy Constraints," in *Proc. IEEE International Conference on Communications (ICC'05)*, 2005, pp. 725–729.
- [21] D. Jin, S. Krishna, R. Kashyapa, and C. Lu Jian, "A multi-layered architecture and protocols for large-scale wireless sensor networks," *Veh. Technol. Conf. 2003. VTC 2003-Fall. 2003 IEEE 58th*, vol. 3, p. 1443–1447 Vol.3, 2003.
- [22] M. Zorzi, "A new contention-based MAC protocol for geographic forwarding in ad hoc and sensor networks," *IEEE Int. Conf. Commun.*, vol. 6, no. c, pp. 3481–3485, 2004.
- [23] R. Rugin and G. Mazzini, "A simple and efficient MAC-routing integrated algorithm for sensor network," *IEEE Int. Conf. Commun. (IEEE Cat. No.04CH37577)*, vol. 0, no. c, p. 3499–3503 Vol.6, 2004.
- [24] A. Roy and N. Sarma, "AEEMAC: Adaptive energy efficient MAC protocol for wireless sensor networks," *2011 Annu. IEEE India Conf.*, pp. 1–6, 2011.
- [25] S. Hayat, N. Javaid, Z. A. Khan, A. Shareef, A. Mahmood, and S. H. Bouk, "Energy Efficient MAC Protocols," pp. 1–8.
- [26] K.-W. Chin and D. Klar, "E2MAC : An energy efficient MAC for RFID-enhanced wireless sensor networks," *Pervasive Mob. Comput.*, vol. 7, no. 2, pp. 241–255, 2011.
- [27] M. M. Bhuiyan, I. Gondal, and J. Kamruzzaman, "I-MAC : Energy Efficient Intelligent MAC Protocol for Wireless Sensor Networks," 2011 17th Asia-Pacific Conf. Commun. 2nd – 5th Oct. 2011 | Sutera Harb. Resort, Kota Kinabalu, Sabah, Malaysia, no. October, pp. 133–[1].
- [28] C. Cano, B. Bellalta, J. Barceló, and A. Sfairoupolou, "A novel mAC protocol for event-based wireless sensor networks: Improving the collective QoS," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5546 LNCS, pp. 1–12, 2009.
- [29] N. Javaid, A. Ahmad, A. Rahim, Z. A. Khan, M. Ishfaq, and U. Qasim, "Adaptive medium access control protocol for wireless body area networks," *Int. J. Distrib. Sens. Networks*, 2014.
- [30] S. Escolar, S. Chessa, and J. Carretero, "Energy-neutral networked wireless sensors," *Simul. Model. Pract. Theory*, vol. 43, pp. 1–15, 2014.

AUTHORS PROFILE



Rajan Sharma has received his degrees of B.Tech & M. Tech from Punjab Technical University, Jalandhar, Punjab, and pursuing Ph.D. in the field of Electronics and Communication from the same University. He is having experience of 13 years in teaching. His research areas of interest are wireless communication, computer networks and wireless sensor networks.



Dr. B.S. Sohi has received his degrees of B.Sc. Engineering, Master of Engineering & Ph.D. in Electronics in years 1971, 1981, 1992 from Panjab University, Chandigarh. He is the Ex- Director of UIET, Panjab University, Chandigarh and presently he is working as Professor in the Department of Electronics and Communication Engineering at CGC Group of Colleges, Mohali, Punjab. He is having experience of 35 years in teaching and administration. He has 105 research publications in various fields. His areas of research are wireless networking, computer networking and wireless sensor networks.

Performance Analysis of adaptive routing protocol for cognitive radio wireless sensor networks using bio-inspired methods

¹Amit N. Thakare, ²Dr. Latesh Malik(Bhagat), ³Mrs. Achamma Thomas

^{1,3}Department of Computer Science & Engineering, G.H. Raisoni College of Engineering, Nagpur, MS, India

²Department of Computer Science & Engineering, Government College of Engineering, Nagpur, MS, India

Abstract: Wireless sensor network is the need of today's and next generation requirement to provide the optimal solution for routing techniques in telecommunication. The many different challenges and issues are worked out by various authors. We consider the fault-tolerance issues with multi-sink and multi-channel probabilistic approach. The sensor nodes are deployed in the region in an unattended fashion for data gathering from different places. The swarm optimization techniques as artificial intelligence are used to find out the optimal solution. In the communication network, there is a need for designing a fault-tolerance routing protocol. The node energy depletion as the nodes are in active mode. The node gets exhausted as it is used continuously for routing the packets. We extend our work in designing the fault-tolerance issues using the multi-sink approach as probabilistic techniques inspired from ant colony optimization for cognitive radio wireless sensor network (CR-WSN) using spectrum sensing, sharing techniques applying on AODV and DSR routing protocol. In this paper we try to analyze the adaptive routing protocol with existing bio-inspired AntHocNet routing protocol.

Keywords: fault-tolerance, Ant Colony Optimization approach, CRWSN, Multi-sink, Multichannel, performance analysis.

I. Introduction

In the wireless sensor network (WSN) are having a wide variety of applications to solve the multiple issues in wildlife tracking, military applications, environmental monitoring, traffic surveillance, health care, etc. Like the ad hoc network wireless sensor network to acquire the knowledge of cognitive network using spectrum frequency utilization with 2.4 GHz. The next generation communication network using the cognitive radio sensor network as a self-organized network that consists of a large number of low cost and low powered sensor devices called as sensor nodes with cognitive capabilities. Each sensor node is equipped with sensing unit for capturing, the wireless transceiver is used to capture events and reach to the base station. The analogy to this technique is an ant colony optimization designed by M. Darigo provide the routing protocol as novel approach such as AntNet, AntHocNet, BeeHocNet, BeeSensor etc. are developed to provide optimal solutions. The cognitive radio sensor network is a self-organized network as we prove earlier in previous publications. The primary user (PU) is used the licensed spectrum band for data sensing, data communication, and data process. If the PU is unused the secondary user (SU) is used as unlicensed spectrum band as PU-SU activity. The PU arrival time and departure time is maintaining the proper utilization of spectrum band. As soon as PU departure time processed the SU starts the arrival process as spectrum holes utilization.

Here the network communication layer is considered for fault tolerance issues in wireless communication links. All the sensor nodes with cognitive capabilities, select the neighbor.

A cluster using nearest neighbor techniques and select the one of node select cluster head to forward the data packet from one node to other node using multiple channels and multi-sink approach using the radio interference ranges of sensor nodes.

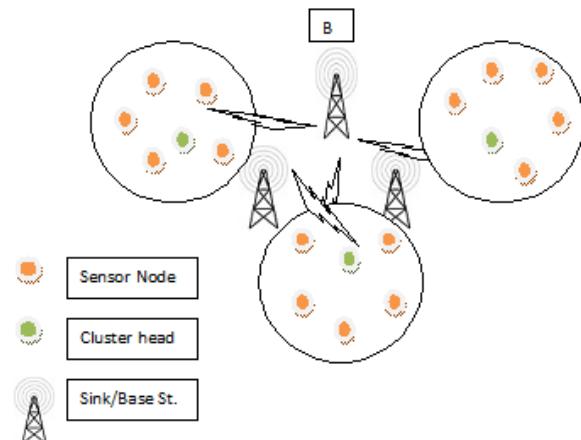


Figure 1. Architecture of multi-sink cognitive radio sensor networks

The retransmission occurs if the failure of data transfer causes a delay, sometimes link failure, and sometimes only using single sink the nodes get exhausted and energy depletion occurs. To solve this problem we are using two sink nodes to avoid the fault occurrences and solve the fault-tolerance issues as similar as inspired ant colony optimization as ANTs starts food searching from source to destination using multiple paths using hops-to-hops to communication as probabilistic approach, either select one path or other as hypothetical approach {0, 1} that it selects either one of the paths or the other path and reach the destination by depositing chemical substances called the pheromone over the all complete path as other ants follow the same path to using pheromone value. The pheromone gets also evaporate as time going on. The obstacle in all complete paths can avoid by ants reach to the destination called forward ants and return back using the same or other bath called backward ants as ACO meta-heuristic.

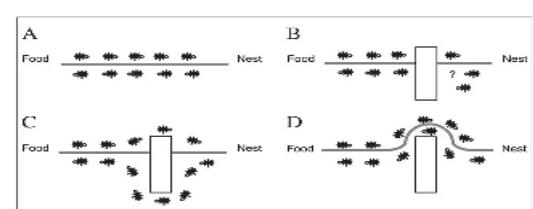


Figure 2. Ant Colony Optimization

In ant colonies, the ant has less intelligence (Caro and Dorigo, 1998) and collective behavior has more intelligence with respect to performing any task during the searching food. They divide the number of the working ants as labor to perform the individual task and this is spectrum sharing in cognitive WSNs.

The many researchers carried out experimentation for WSN and biologically inspired methods. We have gone through the designed protocols for adaptive routing protocols based on WSNs. The next generation novel approach is Cognitive radio sensor networks (E. Fadel, I. F. Akyildiz et al., 2017). Here the utilization of unused spectrum band as unlicensed spectrum band for the other applications, used as to sense the band as spectrum sensing and the spectrum sharing with accessing by the other channel is utilized for the proper flow of data communications between the numbers of users exchanging from node to node and collectively provide to base station. Thus, the analogies between the cognitive radio sensor networks as with Ant Colony Optimization (ACO) provides better outcomes for finding the optimized solution for routing techniques in a telecommunication network.

The rest of the section is as follows, Section 2 describes the literature review; section 3 design issues; section 4 provides the proposed protocol; section 5 observed the Performance analysis; section 6 Concluded the paper.

II. Literature Review

Hai Liu et al.,[2009] proposed a how fault-tolerance addresses in different applications of WSN. They consider the five different parameters that are node replacement of WSN, topology control, target & event detection, data gathering and aggregation. N.Fami-Tafreshi et al., [2016] proposed the idea of exploiting the mobility of data collection points (sinks) for an increasing network lifetime of WSN. The proposed method is based on ant colony optimization algorithm. Lei Cao, et al.,[2010] proposed multiple sink cluster WSN schemes which combine two solutions, also proposed an efficient transmission power control scheme for sink centric cluster routing protocol in multiple sink WSN that is distributed, scalable, self-organized, adaptive system. Athar Ali Khan et al., [2017] evaluate the requirement and key design challenges for routing and MAC protocols in the CR-based smart grid & provide research and review of research carried out for CR based routing. E. Fadal et al., [2017] proposed honey bee mating optimization based routing and cooperative channel assignment algorithms. The development of framework decreases the probability of packet loss and preserve high link quality among sensor nodes in the harsh smart grid environment and performance evaluations is done based on the packet delivery ratio, transmission delay, and residual energy. A. Ozan Bicen, Ozgur B. Akan [2011] point out the need for a novel reliability and congestion control mechanism for CRSN with transport layer simulated & demonstrate performance of protocol in CRSN and much more.

III. Design issues of routing protocols

Wireless sensor network has different challenges and issues to overcome the constraints of WSN's and solve the design application issues. The challenges and design issues in WSN are limited energy consumptions, fault tolerance, scalability, productive cost, data aggregation, load balancing, congestion, security, self-organization.

Self-organization: A WSN is expected to remain operational for an extended period of time. Here the new node added in this network, maybe the other nodes fail because of failures or exhaust their batteries becomes un-operational. A routing protocol must resilient to such dynamic & generally unpredictable variations sustain the long-term availability of network services.

Fault-tolerance: CR-WSNs should have self-forming, self-configuration and self-healing properties. In other words, whenever some nodes or links fail, an alternative path that avoids the faulty node or link must be derived. In CR-WSNs, faults can occur for a variety of reasons, such as hardware or software malfunctioning, or natural calamities, e.g., fire, floods, earthquakes, volcanic eruptions, or tsunamis etc. A CR-WSN should always be prepared to deal with such situations. The fault tolerance or reliability of a wireless sensor node using the Poisson distribution within the time interval (0,t) occurs.

Scalability: Sensor node deployment in WSNs is application dependent and affects the performance of the routing protocol. A large number of nodes is deployed in the region having short communication range and high failure rates. The routing protocols have effectively acceptable for those challenges.

IV. Proposed adaptive routing protocol

The objective of the proposed routing protocol to solve the fault-tolerance issue as adaptive in nature in cognitive radio sensor network, that is the next generation wireless communication. The WSN is equipped with cognitive capabilities of spectrum sensing, spectrum sharing, spectrum decision with low power battery and energy level was slightly exhausting after transmission of data from multiple routes with multiple sink and multiple channels. As the number of channels increases the variation in output parameters are reflected in the evolution. The neighboring nodes discover the route and coordinate to transmission and sharing of data collectively and form the cluster. After forming the cluster as per the energy level they all node in cluster chose the cluster head.

The all techniques are inspired from ANT colony optimization for finding the food source by trailing the pheromone. Which was followed further by rest of ANTs to find the shortest route and food source called forward ANTs and reverse came back to their nest. The hurdles in between the route are avoided and find out the alternate path to reach the destination. As they also use multiple paths shortest route to achieve the target. Inspired from biological organism, we try to develop the routing protocol with solving some issues in communication technology. The Pseudo code is given below.

If a sink node fails to send the data to the base station, the solution for improvement for solving the issues are using multi-sink, multichannel operation in network. Ant –based clustering is a probabilistic approach with multi-hops techniques. We introduced such a probabilistic based routing protocol for fault-tolerance issues. [Awwad et al., 2010] Here if the one of the sink node fails the solution is to use multi-sink base station. The process of working in the figure given below that there is deployment of sensor nodes in the region and then started the neighboring techniques; the nodes started finding the neighbor in the probabilistic hops manner. Then there is cluster is formed and chose the cluster head (CH) as per energy level of individual node. As per bio-inspired techniques the Poisson distribution, the arrival and departure rate of PU (primary user) provides the performance analysis and solves the issues.

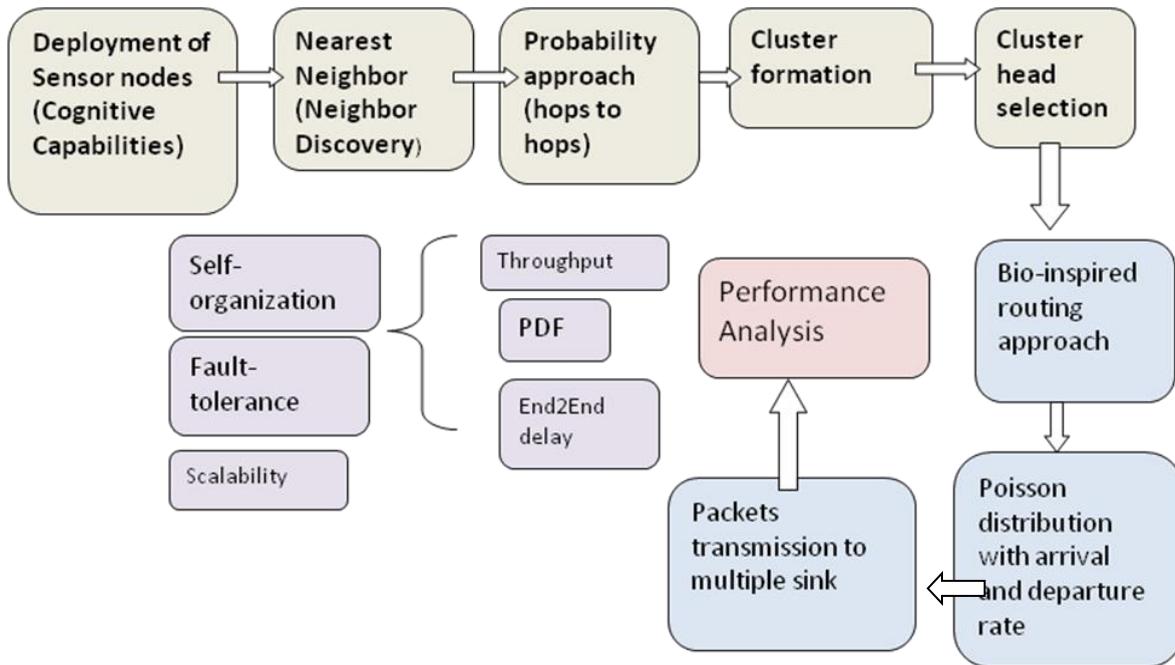


Figure 3. Block diagram of actual working of proposed work

4.2 Pseudo code for routing in WSN

Let consider the CRSN sensor node i and j as edge (i, j) ant start searching food from the best path.

Cluster Head Selection Algorithm

Setup Phase

```

for each path  $(i, j)$  do
    perform spectrum sensing;
end for
finding each hops from  $i$  to  $j$  as best path (shortest)
for each ant  $k_1$  &  $k_2$  do
    Compute probability for choosing multi- path;
    if [node  $j$  nearest neighbor to node  $I$ , ant  $k_1$  &  $k_2$  checks both
    path & take decision  $J \leftarrow$  chose path, then
        Calculate length ;
        Compute as shortest path and form a cluster (all nearer nodes);
    else longest path
        ant  $k$  computes best path  $(i, j)$  by dropping pheromone trail;
    end
    pheromone value = 1;
    if node with greater pheromone value then others then
        ant  $k$  select path  $(i, j)$  in multi-hop fashion;
    else if node which have deeper pheromone value choose as CH
    then
        Advertise as cluster head (CH);
    End
    Transfer data from CH to Master CH (as sink node);
    for each path  $(i, j)$  do
        If one sink exhaust (because of energy level) the other channel
        and sink node chosen
    End

```

Steady Phase: Multi –Hop routing algorithm

```

for each edge do
    Set initial pheromone value 1;
end for
for each CH do compute the visibility
end for
While not stop do
    for each ant  $k$  do
        Select CH;

```

```

for i=1 to n do
    Compute probability  $P_{ij}$  and select next CH  $j$  with;
    Probability  $p_{ij}$ ;
end for
end for
for each path do
    Update the pheromone value;
end for
end while
select shortest path  $ij$  &
Choose the CH as a leader for each CH to base station
do
    sends packet under leader control;
end for

```

Communication inside cluster

```

for each cluster do
    All nodes sends data to CH
    CH check the availability of network
    if (Network is busy ) then
        perform spectrum sensing;
    if a spectrum band is available then
        utilize primary network;
        transfer the data through primary networks(Pus);
        wait for selection of CH and master cluster Head(in which all
        CH sends data);
    if CH is elected in transferring data then
        send the data with max packet size;
    else
        send the data with restricted packet size;
    end
    else
        utilize secondary network(SUs);
        Here access SU over CSMA protocol;
    end
Utilize the secondary network;
end

```

In this paper, we consider the three conditions as channel selection, mobility as a static sink. Initially, we consider the static sink.

The initial assumption in the proposed protocol is considered in the following form.

1. All sensor nodes and sink node (Base Station) are stationary after deployment.
2. The sensor nodes are uniformly distributed with random deployment.
3. The sensor nodes with heterogeneous configuration.
4. All nodes have limited energy.
5. Cross layer MAC layer and Network layer jointly work with TCP transport protocol.

In proposing work we consider two methods for our demonstration of the adaptive routing protocol for solving the fault-tolerance issues in communication.

- a. Single sink and multiple channels.
- b. Multiple sink and multiple channels.

Now consider multiple sinks, multiple channels and demonstrate and find out performance. Here while finding the shortest path among routing in communication the wireless sensor network having the capabilities of cognitive radio spectrum utilization. Here the spectrum senses the available spectrum band search for a certain duration of time for neighboring node and jumps on the vacant band, identify the primarily used band if free serve the channel and if busy, it went back for spectrum sensing. The served channel take a decision and senses the carrier which sends the data to the cognitive sensor node, a user (CR) as forwarding ANT send the information to next coming ants through pheromone trails either send back to carrier sensing for the other vacant band.

The communication protocols are not sufficiently flexible regarding environmental changes. These environmental changes and control on each layer in wireless sensor network architecture operate on widely different timescales.

MAC layer supports for one hops communication where data transmission takes a few milliseconds in most sensor networks. Energy efficiency MAC protocol with sleep scheduling for prolonging for a lifetime is assumed in sensor networks.

Whereas routing layer has deals with topological changes to realize source to destination communications. In static sensor nodes manage network topology by using HELLO message every several tens of seconds.

The timescale operation external control of self-organization is longer than routing layer. It is insufficient to discuss the robustness within one layer. We consider two layer combinations as cross layer approach, MAC and Network layer. We have designed for the fault-tolerance based routing protocol for the condition of channel selection, mobility and node failure.

In MAC layer the sleep control is expected, so that power saving option is successful. The MAC protocol with the sleep control allows the node to sleep in every ten milliseconds. So that each node can communicate with each node only when it's awake. The cycle of sleep control means the minimum unit time of one-hop data transmission. In MAC layer the selection of next hop node when it is in sleep mode the data is to hold for certain period of time. There is condition comes to probabilistic channel selection for communication. The channel selection from spectrum sensing, sharing utilizing the available spectrum band called as cognitive radios in nature, self-organization for sensor node in wireless sensor networks, which is to be analogous to biologically inspired methods, i.e., availability of different paths for ANTs and BEEs for searching the food among different path to reach the destination i.e., sharing resources.

Consider the insect colony as cognitive radios networks. The insects are as a cognitive radio for spectrum utilization. Task allocation is available to channel and task associated stimuli as a permissible power to channel.

- 1) Selecting the channel that has the minimum channel selection probability (appropriate channels) with avoiding the interference at the same time.(select the paths by ants and bees as availability of different paths)
- 2) Manage the number of sensor nodes, which are nearer to each other as a neighboring node forming the groups called clusters (the node that has the fewest radiiuses). Gathering the data as a response to sink node for the HELLO message as a multi-hop communication and accomplish the task of joining regarding to pheromone levels.
- 3) If a sink node fails to send the data to the base station. To improve the problem at that moment the multi-sink operation is useful.

A. Markov Chain model for probabilistic approach

The development in a cognitive radio sensor network based on partially observable on Markov chain decision[19]. In different applications the probabilistic approach plays a vital role in solving the number of problems with optimum solution based on Markov chain model.

Markov Chain is a class of characterizing stochastically the process by Markov property. The meaning is the next process takes depends on the current state and not on the previous state. To prove this for discrete time as given as follows and stated by [23]. Here P is the probability of selecting the behavior of next-hop sensor node and X_t is the condition probability.

$$P[X_{t+1}=x | X_0 = x_0, \dots, X_t] = P[X_{t+1}=x | X_t = x_t]$$

This process called as a Markov Chain model.

The Markov chain model for sensor node is based on (ON/OFF) condition of channel selection of channel by primary user and secondary user, also the probability of next hope sensor is either {0, 1} as shown in the figure given below.

In the following condition state that as equivalent to Ants search food stochastically as random walk probability, same as for channel selection by the secondary user after primary user.

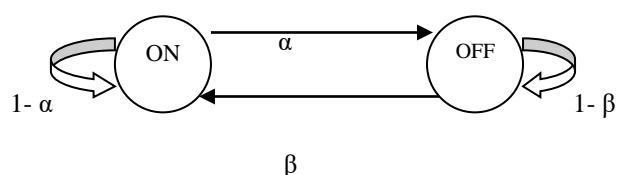


Figure 4. Model for behavior of next-hop sensor node analogous to channel sensing state as ON/OFF.

In random sensing techniques the secondary users cooperate to sense the licensed channels. Each secondary users stochastically chooses a licensed channels for sensing. Su is considered as secondary user as CR node. We can develop the Markov chain model to calculate the probability mass function (PMF) of secondary users and denoted as Probability of secondary user denoted as $\Pr \{Su\}$ for number of secondary users $S = 0, 1, 2, 3, \dots, n$. Each one select n channels as each sensor nodes select the nodes uniformly with 1/n model is

defined for $n + 1$ states. The transition probability from states i to j is $P(i, j)$.

$$Pr(i, j) = Pr(S_{t+1} = j | S_t = i) = \begin{cases} i/n & j = i; \\ 1 - i/n & i = i + 1; \\ 0, & otherwise \end{cases}$$

The above equation, prove that the condition of a number of states from state to another state with probability from one user that is the initial state to another node as the last state with respect to transition from 0 to n and the middle one state is $n-1$ transition with number of probabilities to reach upto j state.

V. Performance results and discussions

To evaluate the performance of the proposed scheme against the recent routing schemes, we considered the three different parameters such as packet delivery fraction, end to end delay and throughput. We evaluated the CRSNs as next generation WSNs. The simulation was done by extending in the open source software NS2 version NS-2.34 with multi-sink & multi-channel extensions to enable dynamic spectrum access and cognitive radio sensor networks. The CRSN network consists of 25 nodes, which are arranged in a grid topology of 1,000m x 1,000m. The licensed arrival and departure time are considered with packet size variation of 100,200,300,400,500 with a queue size of 50. It is assumed that the network utilizes an unlicensed channel which is shared with other legacy networks. This condition causes network interference, which is represented by the parameter (i.e., PU channel). Here we consider the 11 channels for our experimentation in CRSN network. We compare the performance [AODV-Cogns], [DSR-Cogns], AntHocNet. In this simulation, we used Mac/Cogmac standard as the MAC layer protocol. Here we consider the following parameters.

A. Simulation Analysis

The simulation analysis is done on the three parameters, packet delivery factorial, throughput, and delay. We require the high packet delivery factorial, High throughput, low delay. We consider the parameters with their respective values as mention in table 1.

We apply the novel technique on two protocols AODV and DSR with MAC layer protocol as Cognitive radio sensor network. We analyze on two techniques.

1. Packet Variation
2. Pause Time variation
3. Variation on Packet arrival and departure (α, β) respectively.

The packet variation in the novel protocol when applying with the fixed number of nodes deployed in the specific region is given below.

a) Throughput: The parameter evaluation in wireless sensor networks with the considering the above mentioned parameters in NS2 was evaluated. Depends on analysis of throughput with respect to packet size shown in the screenshot above. In the simulation throughput of the network is decreased with respect to increasing the number of packet size. The throughput is calculated as.

Table 1. Parameter Consideration

Simulation Parameters	Values
Channel Type	Channel/WirelessChannel
Radio-propagation model	Propagation/TwoRayGround
Network Interface Type	Phy/WirelessPhy
MAC type	MAC/cogmac/802_11
Interface queue Type	Queue/DropTail/PriQueue
Antenna Model	Antenna/OmniAntenna
Max packet in ifq	50
Number of Mobile nodes	25
Routing Protocol	/AODV/DSR /AntHocNet
Topography	1000mx1000m
Energy Model(Initial)	100 Joule
Data transmission rate	2 sec
Radio transmission range	200
Pause time	0,30,60,90,120,150 sec.
Packet Size	100,200,300,400,500 bytes
Constant Bit Rate (CBR)	256 kb
Channel Number Type/radio	11
Time of simulation	150

Throughput: Total number of bits received /Receiving time of last packet – Sending time of first packet.

b) Average End-to-End Delay: The same as throughput the average end to end delay for the 25 number of nodes with respect to packet size with some statistical values mention in each graphical representation. Here the average end to end delay is less at starting state and increasing as packet size increased.

The average end to end delay = total receiving time of i th We observed that as the variation of packets size though we utilize the multiple sink and multiple channel for shortest route to packets transmission between source to destination in communication by gathering data collectively. Each number of nodes discover the route and by hops-to-hops communication with probabilistic approach. The figure 5 shows that as we increase the packet size the End 2 End delay decreases and some are increases. Though the as compare to AODV(Cogns) , DSR(Cogns) and Anthocnet, the DSR shows less End 2 End Delay as compare to other two, means the DSR with CogNS probabilistic provides less delay for packet transmission, so it is best proposed protocol with cognitive probabilistics which solves and provides high performance for issues such as fault-tolerance.

packet of destination – total sending time of i th packet to source / Total number of received packets by destination nodes.

c) Packet Delivery factorial: The PDF is another parameter in the sensor networks also depends on the total number of sending and receiving packets from any source to destination with respect to time. Here initially it is less as the packets size increases, it also increases.

5.1.1 Packet time Variation

a) Throughput Vs packets size

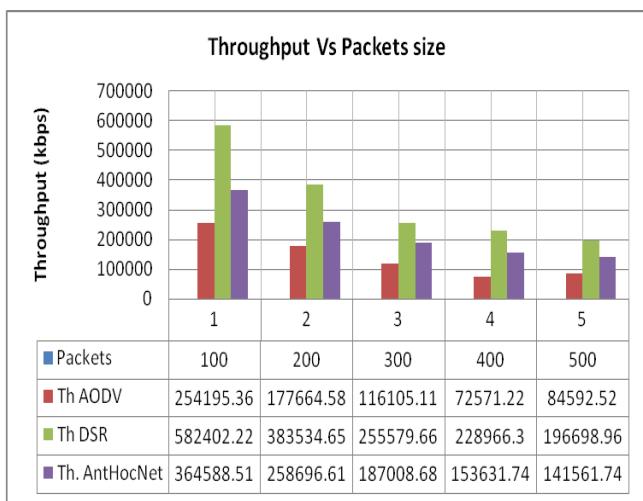


Figure 5. Variation in packets with Throughput analysis

We observed that as the variation of packets size though we utilize the multiple sinks and multiple channel for the shortest route to packets transmission between source to destination in communication by aggregating the data. Each number of nodes discover the route and by hops-to-hops communication with the probabilistic approach. The figure 5 shows that as we increase the packet size the throughput(kbps) decreases. Though the as compared to AODV(Cogns) , DSR(Cogns) and Anthocnet, the DSR shows High Throughput as compare to other two, means the DSR with CogNS capability provides high throughput for packet transmission.

b) End 2 End Delay Vs Packets Size

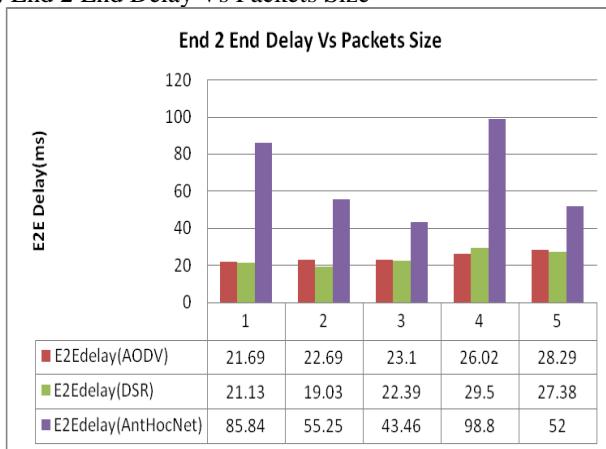


Figure 6. Variation in packets with End 2 End Delay analysis

We observed that as the variation of packets size though we utilize the multiple sink and multiple channel for shortest route to packets transmission between source to destination in communication by gathering data collectively. Each number of nodes discover the route and by hops-to-hops communication with probabilistic approach. The figure 6 shows that as we increase the packet size the End 2 End delay decreases and some are increases. Though the as compare to AODV(Cogns) , DSR(Cogns) and Anthocnet, the DSR shows less End 2 End Delay as compare to other two, means the DSR with CogNS probabilistic provides less delay for packet transmission, so it is best proposed protocol with cognitive probabilistics which⁴⁶

solves and provides high performance for issues such as fault-tolerance.

a) PDF Vs Packets Size

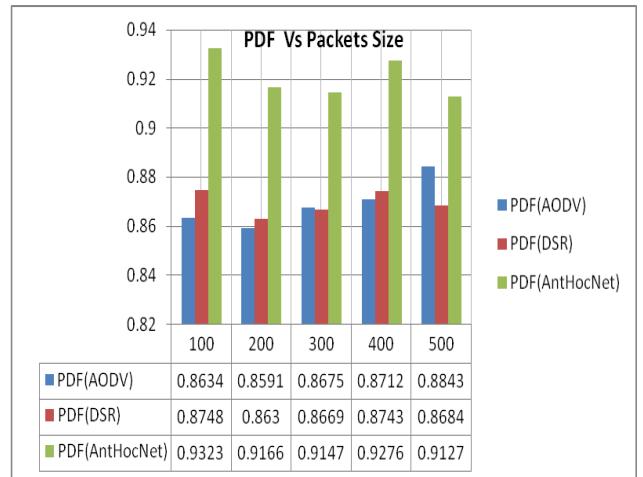


Figure 7. Variation in packets with End 2 End Delay analysis

The figure 7 shows that as we increase the packet size the PDF(packet delivery fraction) increases. Though the as compare to AODV(Cogns) , DSR(Cogns) and Anthocnet, In this case Anthocnet increases in PDF as compare to other two. Means the DSR with CogNS probabilistic still intermediate averagely between AODV and AnthocNet.

Pause Time Variation

a) Throughput Vs Pause Time

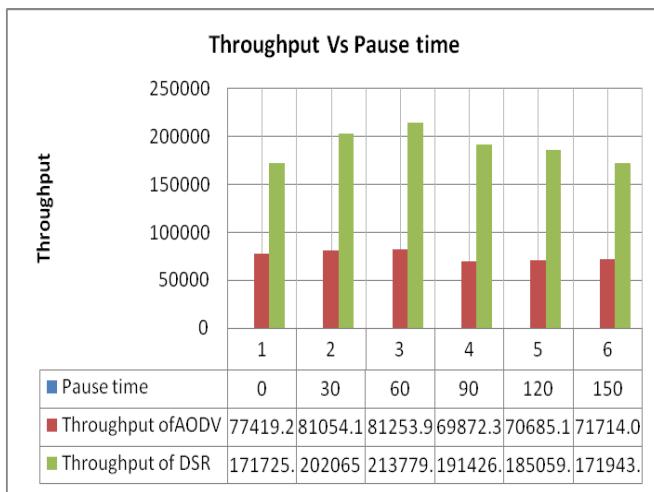


Figure 8. Variation in Pause time with Throughput

The figure 8 shows that as we varies the pause time of 0,30,60,90,120,150. The pause time is defined as the parameter in which the node moves randomly choose path towards destination. When it hits destination pause for some time and changed the direction. where the packet size of 500, it is fixed for all the pausetime. The throughput of DSR with CogNS have high throughput as compare to AODV with CogNS. Here variation in pausetime in AntHocNet never shows any variation.

a) End 2 End Delay Vs Pause Time

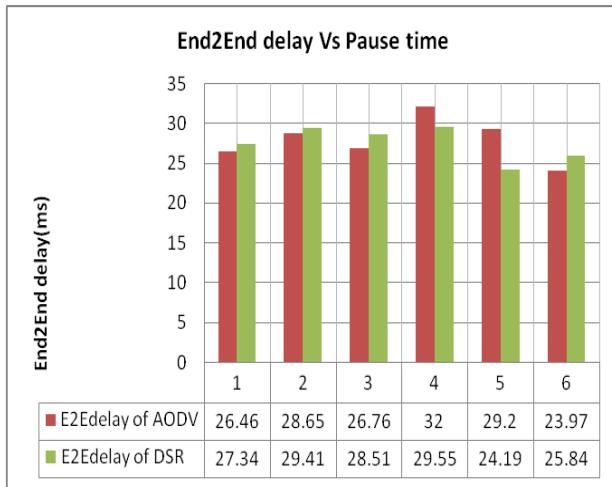


Figure 9. Variation in Pause time with End2End delay

The figure 9 shows that as we varies the pause time of 0,30,60,90,120,150 which we for the packet transmission of 500 in size , fixed for all the pause time. The end2end delay of DSR with CogNS have less delay as compared to AODV with CogNS. Here variation in pause time in AntHocNet do not shows any varaiton.

a) PDF Vs Pause Time

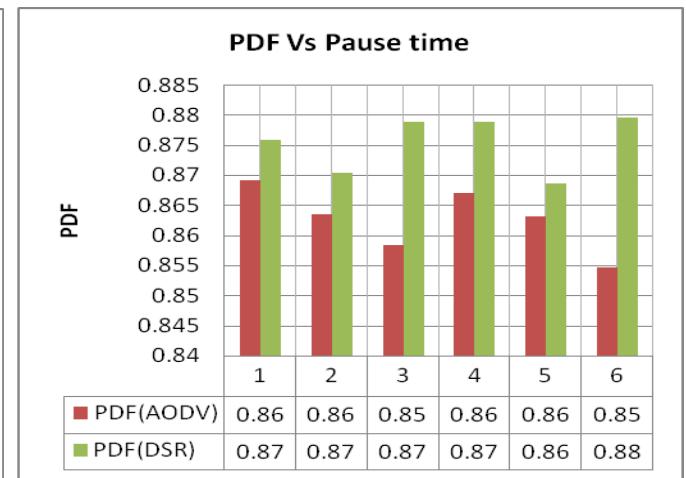


Figure 10. Variation in Pause time with PDF

The figure 10 shows that as we varies the pause time of 0,30,60,90,120,150 which we for the packet transmission of 500 in size , fixed for all the pause time. The end2end delay of DSR with CogNS have high packet delivery ratio as compare to AODV with CogNS. Here variation in pause time in AntHocNet do not shows any varaiton.

5.1.3 Variation scenarios of primary user arrival and departure.

$$\text{Arrival} = \alpha \quad \text{Departure} = \beta$$

- Scenario 1: $\alpha=1, \beta=1$
- Scenario 2: $\alpha=1, \beta=2$
- Scenario 3: $\alpha=1, \beta=3$
- Scenario 4: $\alpha=2, \beta=1$
- Scenario 5: $\alpha=3, \beta=1$

The Primary User has the arrival rate and departure rate applied on different scenarios as mention above. The protocol has multi-sink, multi-channel fashion among different clusters.

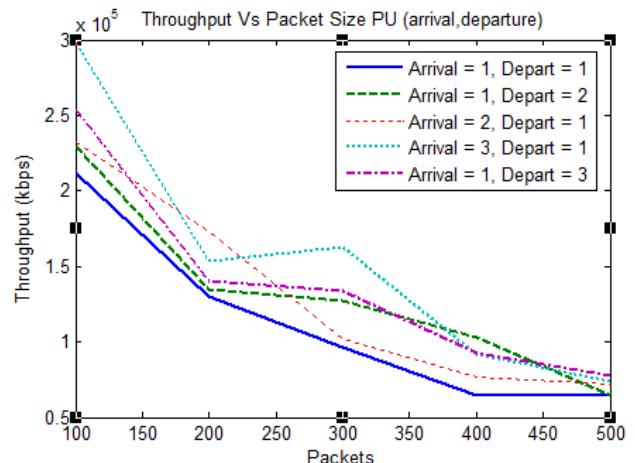


Figure 11. Variation in arrival and departure time with Throughput

Here arrival 3, and departure 1, shows better output of throughput as compare to rest of the transmission of arrival and departure of primary user based on ACO techniques.

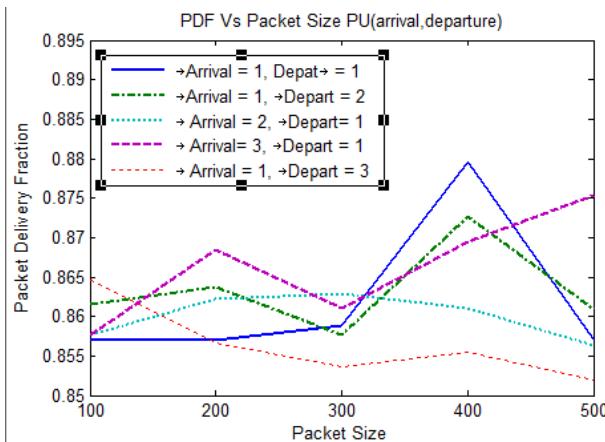


Figure 12. Variation of PU activity with packet delivery fraction

Here arrival 3, and departure 1, shows PDF with high PDF after 300 packet size it gradually increases but arrival with 1, and departure with 1, increases after 300 packet size shows accurate as compare to rest of the transmission of arrival and departure of primary user based on ACO techniques.

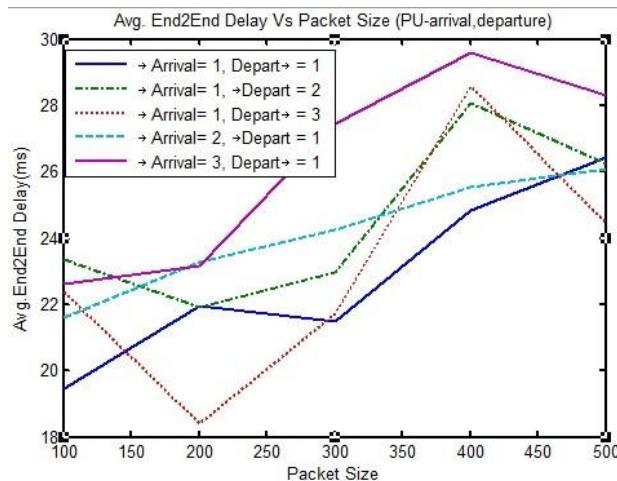


Figure 13. Variation of PU activity with End 2 End delay

Here arrival 3, and departure 1, shows end to end delay after 300 packet size it gradually increases but arrival with 3, and departure with 1, increases after 300 packet size shows accuracy as compare to rest of the transmission of arrival and departure of primary user based on ACO techniques.

Thus the based on three parameters the DSR with Cognitive capabilities based on bio-inspired techniques of Ant Colony optimization provides good accuracy for data transmission. The arrival and departure of PU activity give a spectrum hole for utilization of SU user as a shortest method to route the packet solving the fault-tolerance issue.

VI Conclusion

The proposed bio-inspired cognitive capable routing algorithm with AODV, DSR and AntHocNet routing protocol is

analyzed. It is observed that the proposed method of multi-sinks, multi-channel routing techniques are implemented on AODV and DSR, the findings and observations from above graphs are that the proposed bio-inspired DSR (cogns) gives overall accuracy and reliability with highest throughput and high packet delivery fraction and less end to end delay as compare to AODV (cogns) and AntHocNet with maintaining the energy level. Thus for next generation routing protocol for wireless sensor network which are prone to failure of nodes even though the energy of individual nodes exhaust, finout the alternate solution for the same with using available spectrum bands as secondary users. In future work we have to solve the scalability issues. In future work we solve the scalability. Issue for CRSN.

Acknowledgment

I would like to thanks to my guide Dr. Latesh Malik for their regular guidance and also thanks to all the researcher for their valuable research contribution in my research.

References

- [1] Hang Su, Xi Zhang, "Cross – layer based opportunistic MAC Protocols for QoS Provisioning Over Cognitive Radio Wireless Networks", IEEE Journal on Selecting areas in Communications, Vol. 26, No. 1, January 2008.
- [2] D. G. Reina, S. L. Toral Marin, E. Asimako poulou, F. Barrero, N. Bessis, The role of congestion in probabilistic broadcasting for ubiquitous wireless multi-hop networks through mediation Analysis.
- [3] G. Bolch, S. Greiner, H. de Meer, K. S. Trivedi, and K. S. Trivedi, "Queueing Networks and Markov Chains: Modeling and Performance Evaluation With Computer Science Applications", Wiley-Interscience, 1998.
- [4] V. Esmaelzadeh, R. Berangi, S. M. Sebt, E. S. Hosseini, and M. Parsinia, "CogNS: A Simulation Framework for Cognitive Radio Networks", Wireless Personal Communications, vol. 72, no. 4, pp. 2849 - 2865, Oct. 2013.
- [5] M. D. Felice, K. R. Chowdhury, " Search: A routing protocol for mobile cognitive radio ad-hoc networks", Elsevier , Computer Communications 32(2009) 1983-1997.
- [6] Zhangliang Liang et al., "Delay Performance Analysis for supporting real-time traffic in a cognitive radio sensor network", IEEE Transactions on Wireless Communications, Vol. 10, No. 1, January 2011.
- [7] Ozgur B. Akan et al., "Cognitive Radio Sensor Network", IEEE Network, Vol.23, Issue 4, July-August 2009.
- [8] G. P. Joshi et al., "Cognitive radio wireless sensor networks: applications, challenges and research trends," Sensors, vol. 13, no. 9, pp. 11196-11228, 2013.
- [9] B. Chandrashekaran, "Survey of Network Traffic Models" <http://www.cse.wustl.edu/jain/cse56706/ftp/trafficmodels3.pdf/>
- [10] Le The Dung, et al., "Simulation modeling and analysis of one hop count distribution in cognitive radio ad-hoc networks with shadow fading", Elsevier, simulation, Modeling, practice and theory, 69 (2016), 43-54.
- [11] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "Next generation/dynamic spectrum access/cognitive radio wireless <https://sites.google.com/site/ijcsis/>" ISSN 1947-5500

- networks: A survey,” Computer Networks, vol.50, no. 13, pp. 2127-2159, Sep. 2006.
- [12] Muhammad Saleem, Gianni A. Di Caro, Muddassar Farooq, The survey of Swarm Intelligence based routing protocol for wireless sensor networks, *International Journal of Elsevier on Information Science*, 2010.
- [13] H. Liu, A.Nayak, et al, “Fault-tolerance algorithm/protocols in wireless sensor networks,” in Guide to wireless sensor Networks, S.Misra, Eds. New York: Springer-verlag, 2009, pp. 261-292
- [14] N. Fami-Tefreshi and M Davoudi-Monfared, 2016. Ant Colony Based Mobile Sink Routing Algorithm for Wireless Sensor Network. Asian Journal of information Technology, 15: 2981-2987.
- [15] Cao L, Xu C, Shao W, et al., Distributed Power Allocation for Sink-Centric Clusters in Multiple Sink Wireless Sensor Networks. Sensors (Basel, Switzerland). 2010, 10(3): 2003-2026.doi:10.339/s 100302003.
- [16] Athar Ali Khan, et al., “Cognitive –Radio-based Internet of Things: Applications, Architecture, Spectrum relate Functionalities and future research directions”,
- [17] I. F. Akyildiz et al., “NeXt generation /dynamic spectrum access/cognitive radio wireless networks: A survey,” Computer Network Journal (Elsevier), Sept. 2006.
- [18] Awwad, S.A.B. et al. (2010) ‘Cluster based routing protocol for mobile nodes in wireless sensor network’, 23 May, *Wireless Press Commun.*, Vol. 61, No. 2, pp.251–281.
- [19] Wayne Yang, (2013), “Markov Chains and Ant Colony Optimization”.
- [20] NS2 official website, <http://www.isi.edu/nsnam/ns/>

AUTHORS PROFILE

He is pursuing the PhD.(Computer Science & Engineering)from G. H. Raisoni College of Engineering, Nagpur, M.Tech (Computer Science & Engineering)in 2008 from MGMs College of Engineering,Shree Ramanand Teerth Marathwada University, Nanded, B.E.(Computer Technology) in 2002, from Manoharbhai Patel Institute of Engineering & Technology, Rastrasant Tukadoji Maharaj Nagpur University, Nagpur (Maharashtra), India, He is life member of ISTE, Member of CSI and IEI,Paper presented and Published more than 15 papers.

She has completed Ph.D. (Computer Science & Engineering) from Visvayaraya National Institute of Technology in 2010, M.Tech. (Computer Science & Engineering) from Banasthali Vidyapith, Rajasthan, India and B.E. (Computer Engineering) from University of Rajasthan, India . She is gold medalist in B.E. and M.Tech. She is currently working as Associate Professor & Head of Department in Department of Computer Science & Engineering at Government College of Engineering, Nagpur, MS, India. She has teaching experience of 19 years. Mrs. Latesh Bhagat (Malik) is life member of ISTE, CSI, ACM and presented 46 papers in international journal and 57 papers in international conference. She is recipient of 2 RPS and 1 MODROBs by AICTE. She guided 31 PG projects and 8 Ph.D. students are registered under RTM Nagpur University.

A.Thomas received M.Tech(Computer Science & Engineeirng) in 2013. She is currently the Head of Computer Science Department at G.H.Raisoni College of Engineering, Nagpur. She has completed M.Phil(Computer Science) in the year 2011.Her area of specialization is soft Computing and Data mining.

APPLICATION OF THE ORBAC MODEL IN THE CONTEXT OF PHYSICAL ACCESS CONTROL

Belbergui Chaimaa
STIC Laboratory
Chouaib Doukkali University
El jadida, Morocco
belbergui.c@ucd.ac.ma

Elkamoun Najib
STIC Laboratory
Chouaib Doukkali University
El jadida, Morocco
elkamoun.n@ucd.ac.ma

Rachid Hilal
STIC Laboratory
Chouaib Doukkali University
El jadida, Morocco
hilal.r@ucd.ac.ma

Abstract—Controlling logical access to servers, applications and data is not sufficient in an academic institution such as a faculty. The control of physical access to the premises of this one is also paramount. For the reason that, unauthorized access to the faculty computer room as an example can cause the unavailability of all services. Unfortunately, this aspect of security has almost never been taken into account in research works.

We demonstrate, by the present work, the utility of the use of OrBAC in the context of physical access control, as a solution of all the problems presented above. This, is accomplished through the development of a case study of physical access control within a faculty. The results of tests approve that the defined security policy is dynamic. The permissions of access are activated if the context is enabled, and deactivated if not. During simulation, when the time of the access request is modified or when the location of the access requester change some access rules are deactivated. There are other contexts that were tested.

Keywords-physical access; access control; security policy modeling; OrBAC; MotOrBAC.

I. INTRODUCTION

A faculty consists of premises assigned to a public service, but that are not characterized by the public place nature, it welcomes only specific users. The traditional solutions for controlling physical access to the premises of an institution were based on the use of keys. This method is not secured since it is easy to make a key copy.

The risk increases if the keys are lost or stolen, as this involves access to premises by unauthorized people, and so put the security of the premises at risk.

In the absence of academic solutions to this kind of problems, private companies have taken over and offered physical access solutions using badges.

These solutions offer the advantage of reliability compared to the traditional access management. The loss of badges does not affect the security of the premises since it can be immediately blocked by the administrator. They thus limit the risk of theft and ensure the traceability of events. Simplicity is also a strong point of these solutions resulting in the use of a single badge as a means of access to all premises instead of needing a set of keys.

These solutions also have limits, such as the policy complexity. The definition of access rules is made per person,

not per role, also the redefinition of the whole policy is made if modification of access rules is needed, and the definition of contextual permissions is not possible. Another fault exists, it is the lack of means of; modeling, and testing coherence and conformity to the specifications.

The objective of this work is to control physical access within a faculty as an important pillar of security. This, is done by limiting the access to the authorized persons, in order to avoid the intrusions in the premises of this one.

The solution is based on the technology of badges, which is associated with the use of the formal access control model “Organization based Access Control” (OrBAC) as a remedy for its inadequacies.

The article is organized as follows: Section 2 discusses the overview of the research works dealing with the issue of access control. Section 3 presents the used methods, Section 4 explains the concept of the system and the access control requirements, Section 5 models the access control policy of the faculty with OrBAC, Section 6 presents the simulation and discusses the results, and Section 7 concludes the work.

II. AN OVERVIEW OF THE RESEARCH WORKS DEALING WITH THE ISSUE OF ACCESS CONTROL

The physical access control problematic has almost never been addressed in the research works. Researchers are more interested in logical access within several kinds of systems and networks. Distributed, collaborative and multi-agent systems, web services, social networks, cloud computing and others.

Article [6] provides a language to control access in distributed systems. The article [7] introduces the notion of Team-based Access Control as an approach to collaborative environments. As for paper [8], a new access control model is developed to meet all collaboration requirements. It assumes that users interact with a collaborative application by concurrently editing its data structures. Some solutions have also been proposed for multi-agent systems [9] and [10]. The work [9] describes a typing system for a distributed π -calculus, which guarantees that distributed agents cannot access the resources of a system without first being granted the capability to do so. The paper [10] presents a mobile agent structure which supports authentication, security management and access control for mobile agents.

The paper [11] concerns Web services. It proposes the design of an access control scheme that addresses access control issues, and proposes also an extended, trust-enhanced version of the XML-based Role Based Access Control (X-RBAC) framework that incorporates trust and context into access control.

Several researches are interested in access control for social networks [12], [13] and [14]. The article [12] adopts a rule-based approach for specifying access policies on the resources owned by network participants, with constraints of the type, depth, and trust level of the relationships between nodes. For the article [13], it shows the gap between access control management options offered by Facebook and the real requirements of users in the same context. The paper [14] proposes a new access control solution for Facebook. It is an extension of the model "Organization based Access Control" that is specially dedicated to the social network. This proposal meets the needs of users and covers the deficiencies of the policy currently used by Facebook.

The access control issues in the cloud computing environments were also discussed in many works. The paper [15] proposes an access control model to meet the identified cloud access control requirements. The article [16] proposes a novel distributed architecture embedded principles of computing security. A dynamic access control for securing data in the cloud is also presented in [17]. As for the work [18], it leads to the design of attribute based access control for cloud computing. An onto-ACM (ontology-based access control model) [19] was also proposed as a solution to the difference in service providers and users. Using encryption concept in the context of Cloud is also proposed in [20] and [21]. And finally, the paper [22] proposes a novel access control model for Cloud computing systems inspired by OrBAC model, based on trust evaluation, and introduced a Trusted Third party controlling interactions between access requestor and the system that hosts requested resources.

This observation motivates us to explain the importance of taking physical access control into account in any security policy. Thus, to show the usefulness of modeling the physical access control policy within any organization and the benefits to make it using the OrBAC concepts, such as the dynamism. In this work, the case study is carried out within a faculty.

III. USED METHODS

In this part, the methods used in this work are presented below.

A. Physical access control

1) Definitions

An access control can be logical or physical. It therefore allows to control the access to logical resources such as applications and data, or physical ones like an organization premises.

Physical access control is a technique that consists in subjecting the entrance of a premises, organizations, buildings or places to an access authorization. The aim of this access authorization is to protect persons or properties. The access

control policy may concern, for example, access by staff to sensitive areas such as computer rooms, offices, etc.

2) Physical access control by badges

Access control by badges is a widely used solution. It is used by several kinds of organizations to restrict access, to sensitive premises, to specific persons.

This solution allows not only to manage access permissions to the different premises, but also to keep traceability of all accesses of participants.

The access control platform by badges consists of these elements:

- A main console: Generally, it is a computer configured to control the elements of the badge reader device, to analyze and to keep track of the different accesses.
- Applications: These are applications that allow to define access rules such as the configuration of the door opening. It is also the dashboards allowing the supervision and the traceability of the events collected thanks to the badge reader. Using these applications allow to manage users and control units, and centralize user-related items such as access rights, alarms, and more.
- The badge reader: This is a device used for authenticating an individual requesting access to a given premises.

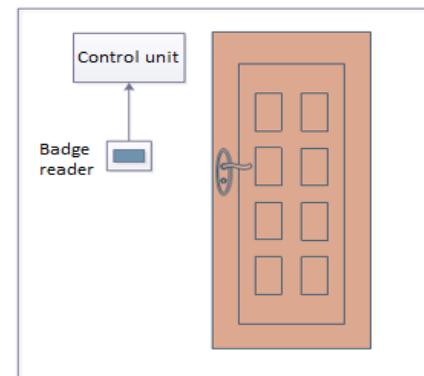


Figure 1. The access control solution by badges



Figure 2. The principal

B. Organization Based Access Control

1) Concept

Traditional access control models are based on the three entities: subject, action, and object. Therefore, access permission specifies whether a subject (s) is authorized to perform an action (a) on an object (o). While, OrBAC adds three abstract entities; role (r), activity (a), and view (v).

Access rules are specified between those entities at the abstract level, and then the concrete access rules can be derived from these ones [3]. The advantage is the independence of the implementation. We present below the abstract and concrete entities, as well as the relations between them, and the definition of contexts.

a) The subject and role entities, and the relationship Empower ()

The role corresponds to the mission of the subject within the organization. Thus, the concept of role models how the organization (org) empowers a subject [4]. Thanks to this concept, the addition of a new subject does not require the modification of the whole policy. It is enough to empower the new subject in the corresponding role since the access rights are already defined.

The relationship between the subject (s) and the role (r) is defined as: Empower (org, s, r). For example, Empower (faculty, Chloe, professor) means that the faculty empowers the subject Chloe in the role professor.

b) The object and view entities, and the relationship Use ()

A view allows modeling and structuring objects. It characterizes how objects are used in the organization (org), and corresponds to a set of objects that satisfy the same property.

The relationship between object (o) and view (v) [4] is defined as follows: Use (org, o, v). For example, Use (faculty, file.doc, administrative document) means that the faculty uses the object file.doc in the view administrative document.

c) The action and activity entities, and the relation Consider ()

The concept of activity allows to model in what way the organization performs actions. The same activity can be implemented in different ways in the same organization [4].

The relationship between action (a) and activity (a) is defined as follows: Consider (org, a, a). For example, Consider (faculty, read, consult) means that the faculty considers the action read as the activity consult.

d) The context and the relationship Define ()

The context is an essential element of OrBAC. The context represents an access constraint [4] which can have different forms; Temporal, spatial, etc. When the constraint is validated, the subject (s) will be allowed to perform the action (a) on the object (o).

The definition of the context is done at the concrete level as follows: Define (org, s, a, o, context). Consider the example: Define (Faculty, Chloe, read, file.doc, working hours). This means that the professor can access the administrative document only during working hours.

2) Expression of security policies in the OrBAC model

Thanks to OrBAC, it is possible to express access rules of different types; Permissions, prohibitions, obligations and recommendations. The notation of an access rule at the abstract level is as follows:

Permission (org, r, a, v, c) means that the organization (org) allows the role (r) to perform the activity (a) on the view (v) if the context (c) is validated.

The notation of the access permission at the concrete level is as follows: Is_permitted (s, a, o). It means that the subject (s) is allowed to perform the action (a) on the object (o). The derivation of access rights at the concrete level is done automatically after defining the relationships between abstract and concrete entities, the context and the access rules at the abstract level as follows:

$\text{org} \in \text{Organizations}, \text{s} \in \text{Subjects}, \alpha \in \text{Actions}, \text{o} \in \text{Objects}, \text{r} \in \text{Roles}, \text{a} \in \text{Activities}, \text{v} \in \text{Views}, \text{c} \in \text{Contexts},$

Permission (org, r, a, v, c) \wedge
 Empower (org, s, r) \wedge
 Consider (org, α , a) \wedge
 Use (org, o, v) \wedge
 Define (org, s, α , o, c)
 \rightarrow Is_permitted (s, α , o)

The notation of prohibitions, obligations and recommendations at the abstract level, as well as the transition to the concrete level are similar to permission definitions.

3) The MotOrBAC simulator

There is a security policy design tool developed with OrBAC. The tool MotOrBAC [5] allows to simulate the access control policy and to test its proper functioning; by checking incoherencies and proposing solutions to resolve it.

IV. CONCEPT AND ACCESS REQUIREMENTS

The concept and the requirements of physical access control within the faculty are presented in what follows.

A. Concept

Every faculty receives a significant number of people per day, they are mostly students, but also professors, researchers, administrative and technical staff, as well as many visitors.

Since intrusions originate in most cases from the interior, and following the problems of disappearance of various objects, it is necessary to restrict the access to the premises (administration, departments, Amphis, etc.) to the authorized persons, by means of automatic opening and closing doors. The opening of each of these doors is controlled by a badge reader placed in the proximity.

Badges are issued to persons who must access the protected premises. Access rules should be defined in such a way as to allow these persons to carry out their tasks within the faculty.

The access rules can be defined for a single door or for a group of doors. Similarly, it may be assigned to one person or to a group of persons. Similarly, a person may have more than a

role, so his access rights should correspond to the union of these roles' access rights.

During this work, we are interested in controlling physical access within a faculty. The premises of the faculty are shown in the figure below.

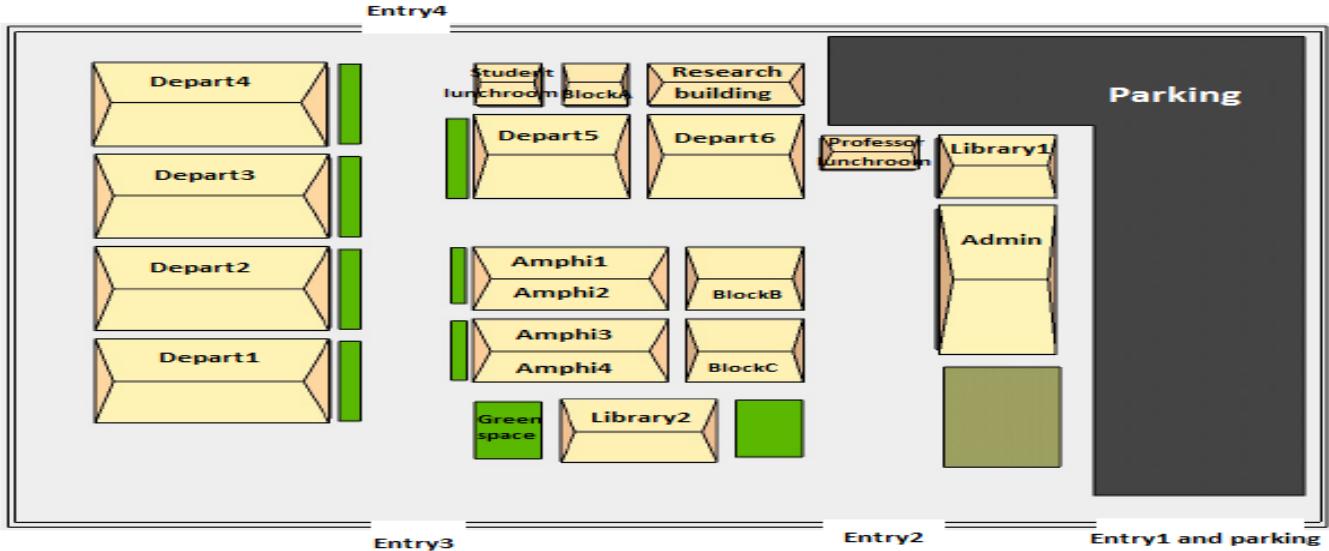


Figure 3. The premises of the faculty

B. Access control requirements

The definition of access rules is made by describing for each person or group of persons the entries authorized to access. Access rules must depend on different constraints; temporal, spatial, and other ones. It is specified in an annual calendar which describes the access requirements, week by week, taking into consideration weekends and holidays. Below is a summary of the recommended access policy.

There are two types of entries to the faculty; an entry that is reserved for professors and staff (Door_for_professors), and three other entries for students and everyone (Doors_for_students). All entries are accessible to professors, staff, and researchers, regardless of the context, even during weekends and holidays. However, students and visitors can access only through the three doors (Doors_for_students), and during working days and hours. The parking is reserved for the exclusive use of professors and staff regardless of the context.

At the administration, administrative staff (Administrative_staff) can access the administration door (Admin_door), has access to the administration offices (Admin_offices) and cannot access the computer room (computer_room), by default. The technical staff (Technical_staff) can access the administration door and has access to the computer room, by default. On the other hand, he can access the administration offices and all the buildings, rooms, offices and amphis only during working days and hours or in case of emergency.

Housekeepers are allowed to access all the buildings of the institution through the main doors during working days, between 6 and 8 am or throughout the day, while access to the offices,

rooms and amphis may be authorized only between 6 and 8 hours.

The remainder of participants can access the administration door only during working days and hours, and the administration offices in the same context but with the presence of the administrative staff. As for access to the computer room, it is strictly prohibited, by default.

In departments, the department door (Depart_door) is always open during working days and hours, and therefore can be accessed by everyone. Heads and secretaries of departments, professors and researchers are allowed to access it also outside working days and hours in case of belonging to the department.

The office of the head of department (office_head_depart) is accessible by him whatever the context. On the other hand, it is accessible by the other persons only during working days and hours with his presence in the office. The same principle is applied to the secretary of department office (office_secret_depart) and to the professors' offices (office_prof_course1, etc.).

Access to classrooms, tutorial rooms (T_room), practical work rooms (PW_room) and research laboratories (researchLab_depart) is prohibited by default for administrative staff and visitors.

However, it is authorized for heads and secretaries of departments, during the working days and hours with the constraint of belonging to the department.

Professors are allowed to access the classrooms, the tutorial rooms and the practical work rooms respectively, during the time, of course, of tutorial or of practical work. As well as for

students in the same context, but with the presence of the professor.

Research laboratories are accessible to researchers, but with the constraint of belonging to the department.

Any Amphi is accessible in two cases; or when an event is organized in it, the constraints are the participation in the event and the time of the event. Either, by the professor, of course, at the course time, and by researchers and students in the same context with the presence of the professor.

Access to tutorial blocks (Block) through the main door is allowed to everyone during working days and hours. On the other hand, access to the block rooms is only allowed to professors of tutorials during the time of tutorials, as well as to students in the same context with the presence of the professor.

The research building is accessible by everyone during working days and hours. However, the laboratories are only accessible by researchers.

Libraries are accessible to everyone during working days and hours.

Two lunchrooms exist; that of the students, but whose access is allowed at the same time to everyone. As well as that reserved for the exclusive use of professors and staff (including the heads and secretaries of departments) and whose access is prohibited for any other person.

Access permissions to classrooms, tutorial and practical work rooms, as well as to amphitheaters, will be made according to the timetable presented in TABLE I.

TABLE I. TIMETABLE OF THE WEEK

	08h	09h	10h	11h	12h	13h	14h	15h	16h	17h	18h
Monday 15 April	Course 1 classroom 1		Course 2 classroom 2				Tutorial 1 Block A room 1	Course 3 classroom 3			
Tuesday 16 April		Course 4 classroom 4		Course 5 Amphi B			Tutorial 2 Block A room 1				
Wednesday 17 April	Course 1 classroom 1		Conference Amphi B				Course 6 Amphi A	Tutorial 3 Tutorial room 1			
Thursday 18 April	PW 1 PW room1				Holiday		PW 2 PW room2				
Friday 19 April	PW 3 PW room3		Course 5 Amphi B				Course 2 classroom 2	Tutorial 4 Tutorial room 2			

V. MODELING OF THE STUDIED SYSTEM WITH ORBAC

The modeling of the studied system with OrBAC is made by firstly specifying the organization that needs to secure access to its premises. After that, abstract and concrete entities, also the contexts should be defined. Finally, access rules can be developed.

A. The organization

In this case study, the physical access control policy will be modelled for a faculty. In the OrBAC sense, the central entity is the organization “faculty” (Fac), bringing together a group of professors, students, etc.

B. Roles and subjects

The faculty (Fac) welcomes researchers, as well as students so that professors teach them collectively. Professors can provide them theoretical but also practical courses, so there are three types of professors; professors of courses, professors of tutorials and professors of practical works. It also welcomes the administrative, technical, and housekeeping staff, one head and one secretary per department, and others. As well as visitors who may be students' parents, information seekers, etc.

Roles and subjects are presented below (Fig. 4).

The faculty empowers subjects in the various roles. Jack in the role of Dean, Tracy and Amber in the role of administrative

staff, and others. The relationship between these subjects and roles is modeled according to OrBAC as follows:

- Empower (Fac, Jack, dean),
- Empower (Fac, Tracy, administrative_staff),
- Empower (Fac, Amber, administrative_staff), etc.

In addition, we use three functions:

- The function Office (subject): It returns the subject office.
- The function geo_loc (subject): It returns the geographical location of the subject during a time t.
- The function Depart (subject): It returns the department to which the subject belongs.

C. Views and objects

The spaces to be protected are divided into several buildings; The administration building, a building per department (there are six departments: Biology, Geology, Chemistry, Maths, Computer Science and Physics), one building per amphi (there are four amphitheaters), one building per tutorial block (there are three tutorial blocks), one research building, one building per library (there are two libraries), one building per lunchroom (there are two lunchrooms, one for professor and one for students).

The entrances to the institution must also be protected (four entries exist, one for professors and staff, and staff and three for all persons). The parking is reserved for professors and staff and must also be protected from unauthorized access.

In the administration, access from the administration door, offices and computer room must be controlled.

Within the department, access from the main door, access to the offices of the head and secretary of the department, the professors' offices, the classrooms, the Tutorial and the practical work rooms, the Research laboratory and the detergent storage room should be restrained.

By space constraints, we cannot enter in the details of each view, some examples are presented (Fig. 6).

The faculty (Fac) considers several objects belonging to the different views which was presented above. The relationship between these views and objects is modeled according to OrBAC as below. The Fac considers the entry1 as an entrance reserved for professors and staff. On the other hand, entries 2, 3 and 4 are accessible by students as well as by other people.

- Use (Fac, entry1, door_for_professors),
- Use (Fac, entry2, door_for_student),
- Use (Fac, entry3, door_for_student) and
- Use (Fac, entry4, door_for_student).

D. The activity “access” and the action “enter”

The main and unique activity in our case study is the “access” to the various premises of faculty. This activity corresponds to the action “enter”. The relation between the activity and action is modeled according to OrBAC as follows:

- Consider (Fac, Access, Enter).

E. The contexts

In this physical access management, the modeling with OrBAC allowed to define the access constraints expressed in the section “requirements”. We have specified several contexts (spatial, temporal, etc.) that can be cumulated according to the security requirements.

The definition of contexts is done according to the timetable (TABLE I).

I) Temporal contexts

All temporal needed contexts are defined below.

a) The default context, it is defined as follows:

$\forall s \forall o \forall a (\text{define}(\text{Fac}, s, a, o, \text{default}) \leftrightarrow \text{true})$: The default context is always true between subject s, action a, and object o.

b) Working days, it is defined as follows:

$\forall s \forall o \forall a (\text{define}(\text{Fac}, s, a, o, \text{working_days}) \leftrightarrow ((D \geq 15 \wedge D \leq 17) \vee (D = 19)) \wedge (\text{MO} = 4) \wedge (Y = 2017))$: In Fac, the context is true between s, a, and o, when the day (D) is between 15 and 17 or equal to 19. The month (MO) is April and the year (Y) is 2017.

c) Working hours, it is defined as follows:

$\forall s \forall o \forall a (\text{define}(\text{Fac}, s, a, o, \text{working_hours}) \leftrightarrow (h \geq 7 \wedge h \leq 19))$: In Fac, the context is true between s, a, and o, when the time (h) is between 7 and 19 hours.

d) Time of event, it is defined as follows:

$\forall s \forall o \forall a (\text{define}(\text{Fac}, s, a, o, \text{time_event}) \leftrightarrow (h \geq 10 \wedge h \leq 12) \wedge (D = 17) \wedge (\text{MO} = 4) \wedge (Y = 2017))$: In Fac, the context is true between s, a, and o, when h is between 10 and 12 o'clock on April 17, 2017.

e) Time of courses

- Time of course1, it is defined as follows:

$\forall s \forall o \forall a (\text{define}(\text{Fac}, s, a, o, \text{time_course1}) \leftrightarrow (h \geq 8 \wedge h \leq 10) \wedge ((D = 15) \vee (D = 17)) \wedge (\text{MO} = 4) \wedge (Y = 2017))$: In Fac, the context is true between s, a, and o, when h is between 8 and 10 of the day 15 or 17 April 2017.

Times of courses 2, 3, 4, 5 and 6 are defined in the same way:

- Time of course2

$\forall s \forall o \forall a (\text{define}(\text{Fac}, s, a, o, \text{time_course2}) \leftrightarrow (((h \geq 10 \wedge h \leq 12) \wedge (D = 15)) \vee ((h \geq 14 \wedge h \leq 16) \wedge (D = 19))) \wedge (\text{MO} = 4) \wedge (Y = 2017))$

- Time of course3

$\forall s \forall o \forall a (\text{define}(\text{Fac}, s, a, o, \text{time_course3}) \leftrightarrow ((h \geq 16 \wedge h \leq 18) \wedge (D = 15) \wedge (\text{MO} = 4) \wedge (Y = 2017))$

- Time of course4

$\forall s \forall o \forall a (\text{define}(\text{Fac}, s, a, o, \text{time_course4}) \leftrightarrow ((h \geq 9 \wedge h \leq 11) \wedge (D = 16) \wedge (\text{MO} = 4) \wedge (Y = 2017))$

- Time of course5

$\forall s \forall o \forall a (\text{define}(\text{Fac}, s, a, o, \text{time_course5}) \leftrightarrow ((h \geq 11 \wedge h \leq 12) \wedge ((D = 16) \vee (D = 19)) \wedge (\text{MO} = 4) \wedge (Y = 2017))$

- Time of course6

$\forall s \forall o \forall a (\text{define}(\text{Fac}, s, a, o, \text{time_course6}) \leftrightarrow ((h \geq 14 \wedge h \leq 16) \wedge (D = 17) \wedge (\text{MO} = 4) \wedge (Y = 2017))$

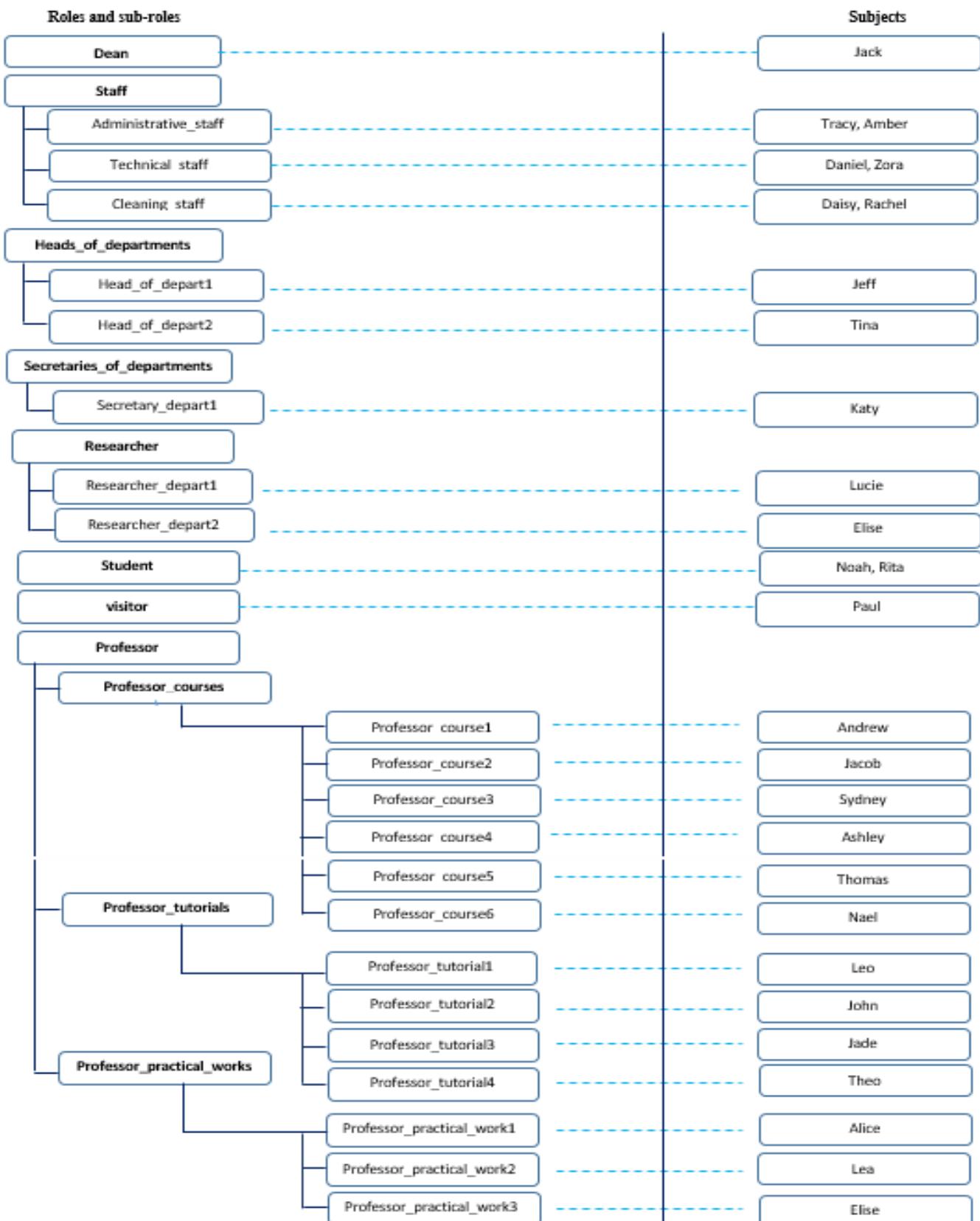


Figure 4. Roles and subjects

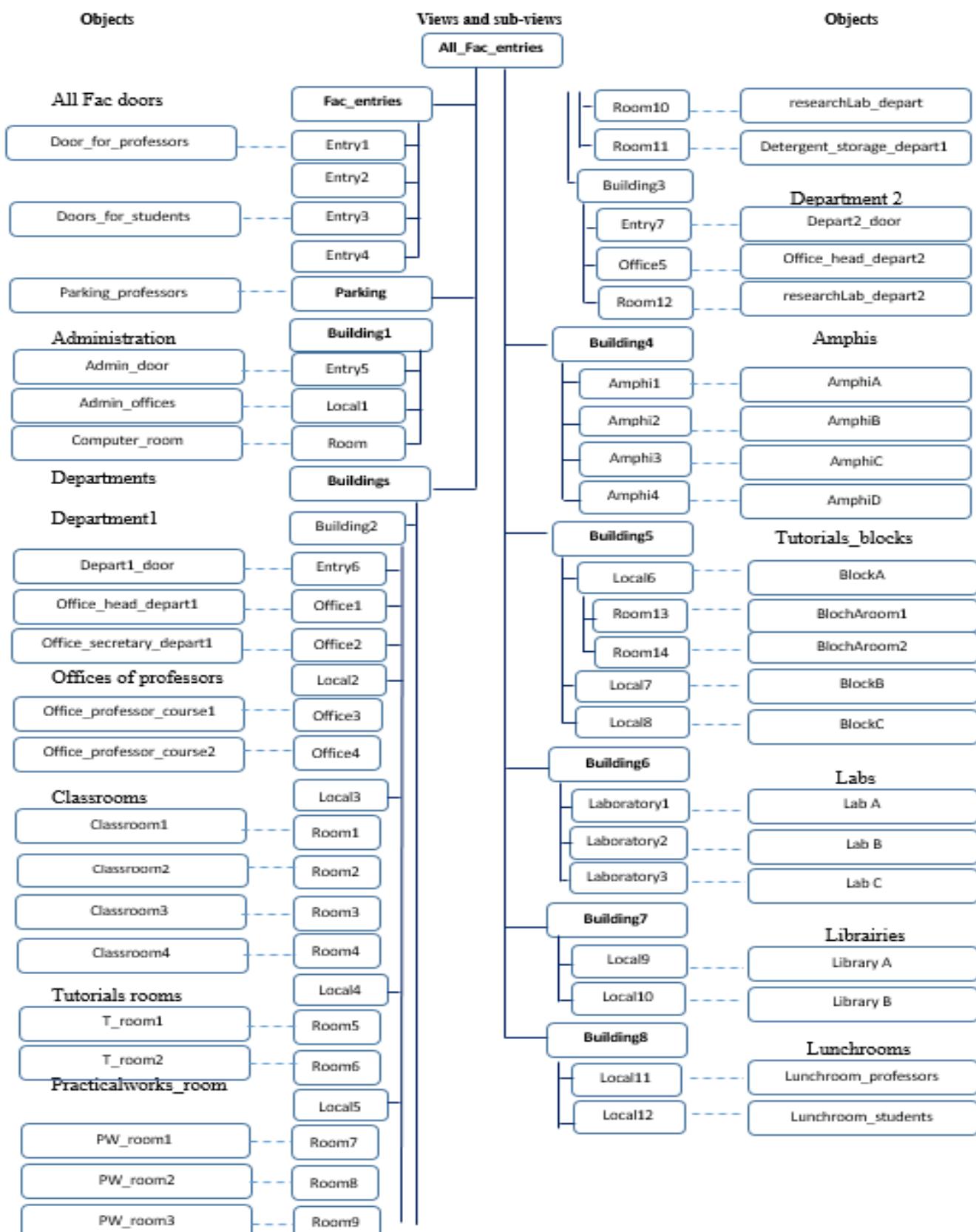


Figure 5. Views and objects

f) Time of tutorials

- Time of tutorial1, it is defined as follows:

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{time_tutorial1}) \leftrightarrow ((h \geq 14 \wedge h \leq 16) \wedge (D=15) \wedge (MO=4) \wedge (Y = 2017)))$: In Fac, the context is true between s, a, and o, when h is between 14 and 16 h of the day April 15, 2017.

Times of tutorials 2, 3 and 4 are defined in the same way:

- Time of tutorial2

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{time_tutorial2}) \leftrightarrow ((h \geq 14 \wedge h \leq 17) \wedge (D=16) \wedge (MO=4) \wedge (Y = 2017)))$

- Time of tutorial3

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{time_tutorial3}) \leftrightarrow ((h \geq 16 \wedge h \leq 18) \wedge (D=17) \wedge (MO=4) \wedge (Y = 2017)))$

- Time of tutorial4

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{time_tutorial4}) \leftrightarrow ((h \geq 16 \wedge h \leq 18) \wedge (D=19) \wedge (MO=4) \wedge (Y = 2017)))$.

g) Time of practical works (PW)

- Time of PW1, it is defined as follows:

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{time_PW1}) \leftrightarrow ((h \geq 8 \wedge h \leq 10) \wedge (D=18) \wedge (MO=4) \wedge (Y = 2017)))$: In Fac, the context is true between s, a, and o, when h is between 8 and 10 of the day April 18, 2017.

The times of PW 2 and 3 are defined in the same way:

- Time of PW2

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{time_PW2}) \leftrightarrow ((h \geq 14 \wedge h \leq 17) \wedge (D=18) \wedge (MO=4) \wedge (Y = 2017)))$

- Time of PW3

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{time_PW3}) \leftrightarrow ((h \geq 8 \wedge h \leq 11) \wedge (D=19) \wedge (MO=4) \wedge (Y = 2017)))$.

2) Spatial contexts

All spatial needed contexts are defined below.

a) Presence

- Presence of administrative staff (pres_admin_staff), it is defined as follows:

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{pres_admin_staff}) \leftrightarrow (\$subject.geo_loc= admin_offices))$: In Fac, the context is true when the administrative staff is present in his office, the geographical location of the administrative staff should be “admin_offices”.

The contexts below are defined in the same way:

- Presence of head of department (pres_resp_depart)
- Presence of secretary (pres.secret)
- Presence of course professor (pres_course_prof)
- Presence of tutorials professor (pres_tuto_prof)
- Presence of practical work professor (pres_pw_prof).

3) The contexts; emergency, belongs to depart and participant

a) Emergency, it is defined as follows:

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{emergency}) \leftrightarrow \text{true})$: In Fac, the emergency context is true when a fault is detected at the equipments or when an alarm is triggered.

b) Belongs to depart, it is defined as follows:

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{Belongs_to_depart}) \leftrightarrow (s \in \text{depart.members}))$: In Fac, the context is true between s, a, and o, when the subject is a member of the department.

c) Participant, it is defined as follows:

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{Participant}) \leftrightarrow (s \in \text{event.participants}))$: In Fac, the context is true between s, a, and o, when the subject appears in the list of the event participants.

4) The composed contexts

Several composed contexts are used in this work:

- Working hours and working days, it is defined as follows:

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{workingdaysANDhours}) \leftrightarrow (\text{working_hours} \wedge \text{working_days}))$: In Fac, the context is true between s, a, and o, when the two contexts working_hours and working_days are true at the same time.

- Time of technical works, it is defined as follows:

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{time_technicalworks}) \leftrightarrow \text{workingdaysANDhours} \vee \text{emergency})$: In Fac, the context is true between s, a, and o, in case of emergency or when the time of the access request respects the context workingdaysANDhours.

- Working hours, working_days and presence of administrative staff

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{workingdaysANDhoursANDpres_admin_staff}) \leftrightarrow \text{workingdaysANDhours} \wedge \text{pres_admin_staff})$: In Fac, the context is true between s, a, and o, when the administrative staff is present in his office and when the time of the access request respects the context workingdaysANDhours.

- Time of course1 and presence of course1 professor

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{time_course1ANDpres_prof_course1}) \leftrightarrow \text{time_course1} \wedge \text{pres_prof_course1})$: In Fac, the context is true between s, a, and o, when the course1 professor is present in the classroom1 and when the time of the access request respects the context time_course1.

Other composed contexts, of time of course and presence of professor of course, time of tutorial and presence of tutorial professor, and time of practical works and presence of practical work professor, are defined in the same way.

- Participant and time of event

$\forall s \forall o \forall a (\text{define } (\text{Fac}, s, a, o, \text{participantANDtime_event}) \leftrightarrow \text{participant} \wedge \text{time_event})$: In Fac, the context is true between s, a, and o, when the access requester is a participant in the event and the time of the access request respects the context time_event.

- Participant and time of event, or time of course5 and presence of course5 prof

$\forall s \forall o \forall a \text{ (define } (\text{Fac}, s, a, o, \text{Participant AND time_event OR time_course5 AND pres_prof_course5}) \leftrightarrow (\text{Participant} \wedge \text{time_event}) \vee (\text{time_course5} \wedge \text{pres_prof_course5}))$: In Fac, the context is true between s, a and o, either when the access requestor is a participant in the event and the time of the access request, respects the time_event context, either when the course5 professor is present in the amphi and when the time of the access request respects the context time_course5.

F. The security policy modeled with OrBAC

The access control policy is defined at two levels, abstract and concrete one.

1) Policy at the abstract level

By space constraint, we will not be able to present all the permissions of policy. We will only present some examples of those that will be tested in the simulation section.

Access permissions are expressed in OrBAC as follows:

- Permission (Fac, technical_staff, access, local2, time_technicalworks)
- Permission (Fac, professor_course1, access, room1, time_of_course1)
- Permission (Fac, student, access, room13, time_of_tutorial2 & pres_prof_tutorial2)

The first permission indicates that the technical staff is permitted to access the local2 during the time of technical works. The second one indicates that even the professor of course1 cannot access classroom1 at any time. Access is only allowed during the time of course1. The third one indicates that the access of the student to the blockAroom1 is only allowed during the time of the tutorial2 with, the presence of the professor of this tutorial.

2) Policy at the concrete level

OrBAC, concrete permissions and prohibitions are derived automatically after defining the relationships between all abstract and concrete entities, the contexts and the permissions at the abstract level. In the example below, the Fac allows technical staff to access local2 during the time of technical work. It empowers Daniel in the role of technical staff, considers the action access as the activity enter, uses the office of course1 professor in local2 and defines the context time of technical works as an access constraint.

This, implies that when the context is verified, the subject Daniel is allowed to enter the office the course1 professor. It is expressed as follows:

Permission (Fac, technical_staff, access, local2, time_technicalworks) \wedge
 Empower (Fac, Daniel, technical_staff) \wedge
 Consider (Fac, enter, access) \wedge
 Use (Fac, office_professor_course1, local2) \wedge
 Define (Fac, Daniel, enter, office_professor_course1, time_technicalworks)
 \rightarrow Is_permitted (Daniel, enter, office_professor_course1).

VI. SIMULATION AND DISCUSSION

We simulate, in this part, the physical access control policy that was modeled in the previous section. We will illustrate the effectiveness of the policy, modeled with OrBAC, through examples.

We will modify the various entries (time of the access request, presence professors, etc.), in each test, in order to verify the effectiveness of the physical access control policy within the faculty, which was modeled by OrBAC in the previous section.

The MotOrBAC software is used in this simulation. The MotOrBAC simulation window allows to view the permissions and prohibitions at the concrete level corresponding to the access rules of the abstract level. These are activated if the context is enabled, and deactivated if not. The color of a permission is green. However, the color of a prohibition is red. When the access rule is activated, the color is clear. If not, it is dark.

Due to the space constraint, we will only present examples of tests. We will try to vary the roles, views, and contexts during the tests.

A. Example 1

The technical staff is permitted to access to the local2 during the time of technical works. The permission at the concrete level is as follows:

- Permission (Fac, technical_staff, access, local2, time_technicalworks)

The “time_technicalworks” is composed context. Its definition is: (working_hours & working_days) | Emergency, with:

- working_hours is equivalent to ((D>=15 & D<=17) | (D=19)) & (MO=4) & (Y = 2017),
- working_days is equivalent to (h>=7 & h<=19),
- Emergency is true if a fault is detected in the faculty equipments or if an alarm is triggered, and false if not.

When either the context (working_hours & working_days) or the emergency context is activated, the permission is granted at the concrete level:

- Is_permitted (Daniel/Zora, enter, office_professor_course1/2)

Test1: The emergency context is false (Fig. 6), but the “workingdays & hours” context is enabled (Fig. 7). Access permission is activated (Fig. 7).

The screenshot shows a table with columns: Contexts, Abstract rules, Concrete rules, Conflicts, and Enti. Under the Abstract rules column, there is a row for 'emergency' with a type of 'user defined c...'. To the right of the table, a modal window titled 'Modify context def' is open, showing the 'Enter new definition:' field with 'Definition' and 'false' selected.

Figure 6. The context emergency is false

The screenshot shows a table with columns: Type, Derives from, Action, Subject, and Object. The table lists several rows of permissions and prohibitions, such as 'per... permission_technical_staff enter Zora office_professor_course2' and 'pro... prohibitionstudent enter Noah PW_room2'. The rows for 'per... permission_technical_staff' and 'pro... prohibitionstudent' are highlighted with red boxes.

Figure 7. Activation of permissions for Daniel and Zora

Test2: The context "emergency" is true (Fig. 8). Therefore, even if the context "workingdays & hours" is not validated because $h = 6$ (Fig. 9), the access permissions are activated (Fig. 9).

The screenshot shows a table with columns: Contexts, Abstract rules, Concrete rule, and Enti. Under the Abstract rules column, there is a row for 'emergency' with a type of 'user defined c...'. To the right of the table, a modal window titled 'Modify context def' is open, showing the 'Enter new definition:' field with 'Definition' and 'true' selected.

Figure 8. The context emergency is true

The screenshot shows a table with columns: Type, Derives from, Subject, Action, and Object. The table lists several rows of permissions and prohibitions, such as 'permi... permission_student4 enter amphiB' and 'prohib... prohibitionstudent enter PW_room2'. The rows for 'permi... permission_student4' and 'prohib... prohibitionstudent' are highlighted with red boxes.

Figure 9. Activation of permissions for Daniel

B. Example

Even the professor of course1 cannot access classroom1 at any time. Access is only allowed during the time of course1. Permission at the abstract level is as follows:

- Permission (Fac, professor_course1, access, room1, time_of_course1), with:

- Time_of_course1 is equivalent to $((h \geq 8 \& h \leq 10) \& ((D=15) | (D=17)) \& (MO=4) \& (Y = 2017))$

When the context time of course1 is validated, permission is activated at the concrete level:

- Is_permitted (Andrew, enter, classroom1)

Test 1: We check the professor's access during the time of course1 (Fig. 10). Access is granted (Fig. 10).

The screenshot shows a table with columns: Type, Derives from, Subject, Action, and Object. The table lists several rows of permissions and prohibitions, such as 'permi... permission_prof_course1 Andrew enter classroom1' and 'prohib... prohibitionresearcher Elise enter PW_room2'. The row for 'permi... permission_prof_course1' is highlighted with a red box.

Figure 10. Activation of permission for Andrew

Test 2: We check the professor's access during non-course1 time (Fig. 11). So, access is not authorized (Fig. 11).

The screenshot shows a table with columns: Type, Derives from, Subject, Action, and Object. The table lists several rows of permissions and prohibitions, such as 'permi... permission_prof_course1 Andrew enter classroom1' and 'prohib... prohibitionresearcher Elise enter PW_room2'. The row for 'permi... permission_prof_course1' is highlighted with a red box.

Figure 11. Deactivation of permission for Andrew

C. Example 3

The access of the student to the blockAroom1 is only allowed during the time of the tutorial2 with, the presence of the professor of this tutorial. Permission at the abstract level is as follows:

- Permission (Fac, student, access, room13, time_of_tutorial2 & pres_prof_tutorial2), with:

- Time_of_tutorial2 is equivalent to ((h>=14 & h<=17) & (D=16) & (MO=4) & (Y = 2017))

- Pres_prof_tutorial2 is true when \$John.geo_loc = BlockAroom1.

When both contexts are activated, access is allowed at the concrete level:

- Is_permitted (Noah, enter, BlockA_room1)

Test1: We show that student Noah cannot access the BlochA_room1 during the non-tutorial2 time (h=13). Permission is disabled (Fig. 12).

Policy Simulation | policy name: physical access control policy

permissions prohibitions obligations

Day: 16 Month: April Year: 2017
Second: 0 Minute: 0 Hour: 13

Type	Derives from	Subject	Action	Object
perm..._permission_professor	Elise	enter	PW_room1	
perm..._permission_student2	Noah	enter	admin_offices	
perm..._permission_prof_course1	Andrew	enter	classroom1	
perm..._permission_visitor	Paul	enter	doors_for_students	
perm..._permission_technical_staff	Daniel	enter	office_professor_course2	
perm..._permission_student3	Noah	enter	amphibB	
perm..._permission_student1	Noah	enter	blockA_room1	
perm..._permission_student4	Noah	enter	amphibB	
perm..._permission_technical_staff	Daniel	enter	office_professor_course1	
perm..._permission_student5	Noah	enter	PW_room2	
perm..._permission_head_depart	Jeff	enter	office_head_depart1	
prohib..._prohibitionstudent	Noah	enter	PW_room2	
prohib..._prohibitionresearcher	Elise	enter	PW_room2	
prohib..._prohibitionresearcher	Elise	enter	PW_room1	
prohib..._prohibitionresearcher	Elise	enter	PW_room3	

Figure 12. Deactivation of permission for Noah

Test2: Even if it is the time of tutorial2, the access of students to the BlockAroom1 is refused if the professor of Tutorial2 (John) is not present (Fig. 13). Permission is disabled (Fig. 14).



Figure 13. Geographical location of John

Policy Simulation | policy name: physical access control policy

permissions prohibitions obligations

Day: 16 Month: April Year: 2017
Second: 0 Minute: 0 Hour: 14

Type	Derives from	Subject	Action	Object
perm..._permission_professor	Elise	enter	PW_room1	
perm..._permission_student2	Noah	enter	admin_offices	
perm..._permission_prof_course1	Andrew	enter	classroom1	
perm..._permission_visitor	Paul	enter	doors_for_students	
perm..._permission_technical_staff	Daniel	enter	office_professor_course2	
perm..._permission_student3	Noah	enter	amphibB	
perm..._permission_student1	Noah	enter	blockA_room1	
perm..._permission_student4	Noah	enter	amphibB	
perm..._permission_technical_staff	Daniel	enter	office_professor_course1	
perm..._permission_student5	Noah	enter	PW_room2	
perm..._permission_head_depart	Jeff	enter	office_head_depart1	
prohib..._prohibitionstudent	Noah	enter	PW_room2	
prohib..._prohibitionresearcher	Elise	enter	PW_room2	
prohib..._prohibitionresearcher	Elise	enter	PW_room1	
prohib..._prohibitionresearcher	Elise	enter	PW_room3	

Figure 14. Deactivation of the permission for Noah

Test3: When access is required at the time of tutorial2 and the professor is present (Fig. 15). The permission is enabled (Fig. 16).



Figure 15. Geographical location of John

Policy Simulation | policy name: physical access control policy

permissions prohibitions obligations

Day: 16 Month: April Year: 2017
Second: 0 Minute: 0 Hour: 14

Type	Derives from	Subject	Action	Object
perm..._permission_student1	Noah	enter	blockA_room1	
perm..._permission_student2	Noah	enter	admin_offices	
perm..._permission_professor	Elise	enter	PW_room1	
perm..._permission_prof_course1	Andrew	enter	classroom1	
perm..._permission_visitor	Paul	enter	doors_for_students	
perm..._permission_technical_staff	Daniel	enter	office_professor_course2	
perm..._permission_student3	Noah	enter	amphibB	
perm..._permission_student4	Noah	enter	amphibB	
perm..._permission_technical_staff	Daniel	enter	office_professor_course1	
perm..._permission_student5	Noah	enter	PW_room2	
perm..._permission_head_depart	Jeff	enter	office_head_depart1	
prohib..._prohibitionstudent	Noah	enter	PW_room2	
prohib..._prohibitionresearcher	Elise	enter	PW_room2	
prohib..._prohibitionresearcher	Elise	enter	PW_room1	
prohib..._prohibitionresearcher	Elise	enter	PW_room3	

Figure 16. Activation of permission for Noah

D. Example 4

A head of department must access the office of the head of department 1 only if he belongs to this department. In other departments, he can access the head department offices only if the head of this department is present in his office. Permission at the abstract level is defined as follows:

- Permission (Fac, head_of-depart, access, office1, belongs_depart1), with:

- Belongs_depart1 is true if \$subject.depart = depart1

When context is enabled, permission at the concrete level is activated:

- Is_permitted (Jeff, enter office_head_depart1)

Test1: The head of department 1 has a permission to access the office of department 1 (Fig. 17) because it is his office (Fig. 18), but access to the office of the head of department 2 is prohibited (Fig. 17) because he does not belong to this department (Fig. 17), and the head of department 2 (Tina) is not present in his office (Fig. 19).

Figure 17. Access rights of Jeff

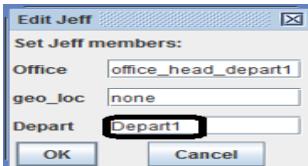


Figure 18. Jeff infos

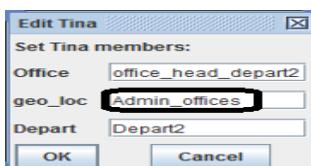


Figure 19. Tina infos

Test2: The head of department 1 has a permission to access the office of the head of department 1 (Fig. 20) because it is his office (Fig. 18). Access to the office of the head of department 2 is also authorized (Fig. 20) because the head of department 2 (Tina) is present in his office (Fig. 21).

Figure 20. Jeff access rights

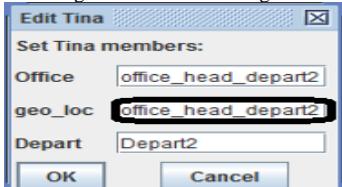


Figure 21. Tina infos

E. Example 5

A visitor can only access the Fac through the doors that are not reserved for professors (entries for student). This access is allowed only during working hours and days. Permission at the abstract level is defined as follows:

- Permission (Fac, visitor, access, entry2, workinghoursANDworkingdays), with:

- workingdays is equivalent to $((D \geq 15 \& D \leq 17) \mid (D = 19)) \& (MO=4) \& (Y = 2017)$ and

- workinghours is equivalent to $(h \geq 7 \& h \leq 19)$

When both contexts are validated, access is allowed:

- Is_permitted (Paul, enter, doors_for_students).

Test 1: Access is allowed because the temporal context is validated (Fig. 22).

Figure 22. The permission is activated for Paul

Test 2: The context "working days" is validated. On the other hand, 20 h does not belong to the time range "working hours", so this context is not validated. The access permission is deactivated (Fig. 23).

Figure 23. The permission is deactivated for Paul

VII. CONCLUSION

In this work, we deal with a serious matter that has almost never been addressed in the research works, to my knowledge, it is the physical access control. We have explained the importance of controlling not just logical access, but also the physical one within any organization. This interest motivated us to model a physical access policy within an academic institution as an example. This, by using the OrBAC model, for all the benefits that it offers; the expression of contextual permissions, the independence of implementation, and others.

We have demonstrated the ease of modeling with OrBAC through excerpts from the access control policy. Subsequently, we have made a test of coherences so as not to have conflicts during the real management of accesses. We have also shown the interest of defining contextual access rules, by modifying the entries; the time of the access request, etc. The access rights assigned to the same role change automatically. Through examples, we have also demonstrated the good functioning of the policy.

ACKNOWLEDGMENT

This research is supported by the National Center for Scientific and Technical Research of Morocco.

REFERENCES

- [1] A. A. E. Kalam et al., "Organization based access control," in Policies for Distributed Systems and Networks, 2003. Proceedings. POLICY 2003. IEEE 4th International Workshop on, 2003, pp. 120–131.
- [2] A. Bâina, "Controle d'accès pour les grandes infrastructures critiques. Application au réseau d'énergie électrique..," INSA de Toulouse, 2009.
- [3] R. Thion, "STRUCTURATION RELATIONNELLE DES POLITIQUES DE CONTRÔLE D'ACCÈS PRÉSENTATION, RAISONNEMENT ET VÉRIFICATION LOGIQUES," Université Paul Sabatier Toulouse, 2008.
- [4] A. A. El Kalam et al., "Or-BAC: un modèle de contrôle d'accès basé sur les organisations," Cahiers francophones de la recherche en sécurité de l'information, vol. 1, pp. 30–43, 2003.
- [5] F. Cuppens, N. Cuppens-Boulahia, & C. Coma, MotOrBAC: "un outil d'administration et de simulation de politiques de sécurité". In Security in Network Architectures (SAR) and Security of Information Systems (SSI), First Joint Conference on June 2006, pp. 6-9.
- [6] M. Abadi, M. Burrows, B. Lampson, and G. Plotkin, "A calculus for access control in distributed systems," ACM Transactions on Programming Languages and Systems (TOPLAS), vol. 15, no. 4, pp. 706–734, 1993.
- [7] R. K. Thomas, "Team-based access control (TMAC): a primitive for applying role-based access controls in collaborative environments," in Proceedings of the second ACM workshop on Role-based access control, 1997, pp. 13–19.
- [8] H. Shen, & P. Dewan, "Access control for collaborative environments", in Proceedings of the 1992 ACM conference on Computer-supported cooperative work (pp. 51-58). ACM.
- [9] M. Hennessy and J. Riely, "Resource Access Control in Systems of Mobile Agents," Information and Computation, vol. 173, no. 1, pp. 82–120, Feb. 2002.
- [10] V. Roth and M. Jalali-Sohi, "Access control and key management for mobile agents," Computers & Graphics, vol. 22, no. 4, pp. 457–461, 1998.
- [11] R. Bhatti, E. Bertino, and A. Ghafoor, "A trust-based context-aware access control model for web-services," in Web Services, 2004. Proceedings. IEEE International Conference on, 2004, pp. 184–191.
- [12] B. Carminati, E. Ferrari, and A. Perego, "Enforcing access control in Web-based social networks," ACM Transactions on Information and System Security, vol. 13, no. 1, pp. 1–38, Oct. 2009.
- [13] C. Belbergui, N. Elkamoun, and R. Hilal, "Modeling Access Control Policy of a Social Network," International Journal of Advanced Computer Science and Applications, vol. 7, no. 6, 2016.
- [14] C. Belbergui, N. Elkamoun, and R. Hilal, "An access control model for a social network: Case of Facebook," International Journal of Computer Science and Information Security (IJCSIS), vol. 14, no. 11, 2016.
- [15] Y. A. Younis, K. Kifayat, and M. Merabti, "An access control model for cloud computing," Journal of Information Security and Applications, vol. 19, no. 1, pp. 45–60, Feb. 2014.
- [16] A. Almutairi, M. Sarfraz, S. Basalamah, W. Aref, and A. Ghafoor, "A distributed access control architecture for cloud computing," IEEE software, vol. 29, no. 2, pp. 36–44, 2012.
- [17] M. Auxilia and K. Raja, "Dynamic Access Control Model for Cloud Computing," in Advanced Computing (ICoAC), 2014 Sixth International Conference on, 2014, pp. 47–56.
- [18] A. R. Khan, "Access control in cloud computing environment," ARPN Journal of Engineering and Applied Sciences, vol. 7, no. 5, pp. 613–615, 2012.
- [19] C. Choi, J. Choi, and P. Kim, "Ontology-based access control model for security policy reasoning in cloud computing," The Journal of Supercomputing, vol. 67, no. 3, pp. 711–722, Mar. 2014.
- [20] S. Ruj, A. Nayak, and I. Stojmenovic, "DACC: Distributed Access Control in Clouds," 2011, pp. 91–98.
- [21] Y. Zhu, H. Hu, G.-J. Ahn, D. Huang, and S. Wang, "Towards temporal access control in cloud computing," in INFOCOM, 2012 Proceedings IEEE, 2012, pp. 2576–2580.
- [22] C. Belbergui, N. Elkamoun, and R. Hilal, "A dynamic Access Control Model for Cloud Computing environments," In Proceedings of the 6th International Conference on Communication and Network Security, 2016, pp. 21-29. ACM.

AUTHORS PROFILE

Belbergui Chaimaa - was born in Morocco in 1991. She obtained her Masters in networks and systems from "Sciences and technologies Faculty of Settat" in 2013. She is currently studying for a doctorate at Chouaib Doukkali University in Morocco. Her field of interest is Modeling and assessment of the security of a companies' information systems.

Najib Elkamoun – received his Ph.D. degree in Optical and Microwave Communication from the National Polytechnic Institute of Grenoble, France, in 1990. He is currently Professor Researcher at Faculty of Science, University Chouaib Doukkali, El Jadida, Morocco. With over 20 years of expertise in information technology and communication, he has conducted several thesis and overseas missions in e-learning and telecommunication networks. His research interests include High Speed Network Architectures (MPLS), Mobility Management, security and QoS in Emerging Networks (MANET, VANET and WSN), Wireless Communications and Traffic Engineering for Computer and Telecommunication Networks.

Rachid Hilal – was born in Rabat, Morocco, in 1965. He received the PhD degree in telecommunications from the University of Limoges, France, in 1996. Since 2003, he was the General Secretary of the Cadi Ayyad University Marrakech. Since 2012, he is a vice president of the Chouaib Doukkali University El Jadida. He is member of the STIC Laboratory and ability professor. His research interests include distributed power amplifiers, microwave nonlinear circuit design, and inset-fed patch antenna.

CLUSTERING ALGORITHMS: A REVIEW

NIKITHA JOHNSIRANI VENKATESAN, CHOONSUNG NAM, DONG RYEOL SHIN*

School of Electrical and Computer Engineering,

Sungkyunkwan University Suwon, South Korea

E-mail: nikithajv@skku.edu, namgun99@gmail.com, drshin@skku.edu

ABSTRACT

In the last few years, large-scale data analysis has become a hot topic in both research and industry side. The traditional database system could no more handle the storage and processing of big data in an individual system. Clustering plays a major role in data analysis. Since 90 % of the total data we deal today are unstructured, classification and recommendation cannot be of great use. Clustering comes into play to deal with a large amount of unstructured data.

Apache Mahout, an open source platform which is focused in of clustering and classification is used on top of Apache Hadoop. In this paper, we study various clustering algorithms on text and XML dataset. We also analyzed the best clustering algorithm by varying important parameters such as distance measure, convergence threshold, and the like. This paper not only gives a detailed survey of apache hadoop and mahout but also presents a case study evaluation of cluster quality. The working of mahout clustering algorithms and Hadoop distributed file system (hdfs) is reviewed in an exhausted manner. We used text and XML dataset of size 7.27 GB (7,811,654,084 bytes) for clustering. Comparison of results in accordance with a distance measure, a clustering algorithm, convergence parameter is provided. This paper helps researchers to analyze the merits and demerits of various clustering algorithms on different kinds of datasets.

KEYWORDS: big data, Apache mahout, clustering algorithms, Hadoop

1. INTRODUCTION

In today's digital era, Big Data is everywhere from social media to sensor data. The rapid evolution of technology is sweeping the entire world [1] with digital information. Humans are connected with each other through a social platform

where neither time nor space act as barriers to communication. Statistics show that it will take 98 years for a person to watch a one day's YouTube videos back-to-back continuously. Clustering [2] is a hot topic of research in different sectors such as management, biology, social media, statistics, and the like. Any kind of objects can be clustered if we can distinguish the objects into similar and dissimilar. For example, Photos and pictures can be clustered depending upon the colors, date of the picture taken, animals or humans, Friends, and families etc.

Today data is being collected almost from everything and everywhere. Our entire life is dependent on data and information. The world has seen exponential growth in data because of social media, health organizations, sensors and so on. Big Data has earned a great significance among the researchers in the past years. "Big Data", as the name depicts, refers to really large data which cannot be handled by the traditional storage system. Hadoop comes into play for storing and processing the Big Data [3]. The data which is stored without any further processing or analyzing is of no use. The main purpose of storing a large amount of data is to find some useful pattern out of it or to predict any fault detection in advance. For example, in health organizations, the data is collected from the patients to predict the various kinds of disease that can occur in future [4]. A famous use case is predicting the cancerous tumor based on its characteristics and features [5]. Similarly, sensor data which are collected from vehicles are of great use to insurance companies [6]. The data analysts have information like driving speed, braking, mileage etc. which are useful to know about the driver. A business organization mainly collects the customer data to understand their churn [7]. After the collection of data, there are many problems in analyzing the data like, finding patterns from the data and identifying new patterns from the incoming datasets. Hadoop provides an open source tool called Mahout which can be very much

useful in finding out the patterns from the datasets. In Machine Learning [8], supervised and unsupervised learning is two broad categories. Supervised learning is further categorized into regression and classification. In the former, the input variables are mapped to some continuous function as output while in the latter, the input is predicted based on its characteristics and mapped into the predefined discrete category as output. Unsupervised learning has little or no idea about what the output should be. Based on the similarities of the input data, the output is predicted and clustered according to the relationships among the data.

In this paper, we will focus on clustering, which is unsupervised learning. We took different types of datasets say text and XML (Extensible Markup Language). The important algorithms are applied to the datasets and the resulting clusters are observed. Similarly, by varying the distance measure and the size of the datasets, the quality of the clusters and the time taken to compute the clusters are noted. Finally, we provide the justification for which algorithm is best suited for various kind of datasets. The organization of the paper is as follows: Section 2 talks about literature review and real-time applications of Mahout Clustering [9]. The working of Apache Mahout and Apache Hadoop is discussed in Section 3. Section 4 exhaustively describes the pseudo code of clustering algorithms step by step. Section 5 and 6 interpret our experiments and results. Finally, Section 7 concludes the paper with possible future work

2. BACKGROUND & LITERATURE REVIEW

The basic idea of clustering is to group the unordered items based on some manner of homogeneity. In technical words, clustering helps to find a proper structure in a collection of unlabeled data. Clustering involves three important things:

- An Algorithm: A step by step procedure used to group items together
- An idea of similarity and dissimilarity: Evaluating which item belongs to which cluster
- A stopping condition: A condition in which the items cannot be grouped further.

The above is the general idea which is being used in various kinds of clustering algorithms based on the size of the data and the results we want. The following section briefs the various research articles that use Apache Mahout clustering algorithms for the respective datasets.

In [10], topic modeling in Mahout has been used for clustering the large amount of data in business organizations. To find out the hidden topics and to eliminate the unwanted noises in Hadoop framework, Latent Dirichlet Allocation is used. A reuters-21578 text corpus is feed into LDA [11] by extending the features of collapsed variational Bayesian (CVB). The results indicate better accuracy in finding the topics among multiple documents when compared to standard CVB. An interesting case study on tourism agency in Sri Lanka has used Apache Mahout collaborative filtering for data analysis [12]. The authors analyzed the data from social media, visitor's blog and the feedback from the tourists. They designed an architecture using Hadoop open source components, where Mahout plays the major role in analyzing the data. Similarly, Arantxa et al. Proposed an open source solution for analyzing the technical call centers data [13]. The architecture includes Mahout as the main components it does the clustering and analysis part. Though other open source components such as Hive and HBase are useful for storage and querying, analyzing the data is the main task to find out patterns from the data. [14] Did a comparison between a variety of data mining frameworks such as weka, rapid miner, Mahout and R. Log files are collected from the virtual campus of Open University of Catalonia and used as a dataset. The dataset is really large and Mahout stood out in performance as well as in clustering quality. Rui Mximo Esteves et al. Clusters all the Wikipedia articles in XML format using K-means and Fuzzy k-means algorithm in Mahout [15]. Since their research was conducted in an early stage, Mahout could not remove the noise in the dataset. The results concluded that the data should undergo lots of preprocessing for a better quality of clustering. Mahout provides the facility for creating our own distance measure as well clustering algorithm [16]. The authors have improved the cosine distance measure and applied the k-means clustering for Wikipedia and Reuters data set. The improvised version of the cosine distance measure shows better accuracy than the standard cosine distance measure. In [17], cloud-based mahout's k-means algorithm is utilized for analyzing data. As healthcare organizations deal with a large amount of data, the authors used healthcare data set for discussing the Mahout clustering. They concluded that Apache Mahout scales well for the enormous amount of data and as the number of nodes gets increased, the algorithm takes less amount of time to cluster. [18] and [19] has focused on social media analyzing. Twitter apparently uses lots of hashtags and they mean the

emotions and moods of the users. Connections between different hashtags have been clustered based upon some sort of similarity. The experiment was done by distributed Hadoop environment using Mahout clustering algorithms. The results revealed great connections between hashtags and topmost words of the cluster. [20] also deals with twitter data clustering based on weighting mechanism and stop words. ChaoLung et al. did real-time crop data analysis in Taiwan using Mahout clustering [21]. The crop dataset contains the species of the crops from the agriculture department. Analysing this data using Mahout and Hadoop increases the seasonal effect in terms of price and volume. Venkateswara Reddy et al. did a comparative study on K-Means, fuzzy and Canopy clustering algorithms at [22]. Two different text data sets taken from UCI repository and Reuters data set are used for the experiments and comparison. The comparison was made between in-memory implementation and MapReduce phase. The results concluded that K-means is not the best suit for Big Data as it suffers in identifying the number of clusters initially. Most of the researchers in their articles use real-world datasets for analysis.⁵

Big organizations such as BuzzLogic, Breast cancer center, stack overflow which is an online discussion website are all using Apache Mahout for analyzing a large amount of data. In medical image processing, Mahout clustering is utilized for image segmentation for CAD (Computer Aided Disease) diagnosis. The mammogram in the breast is clustered into several parts for the future identification of cancer affected region. The real-time applications of Mahout clustering have been exhaustively discussed. From the literature review, it is clear that Mahout is a successful and satisfactory platform for data analysis. The next section gives denotation about the in-depth working of Apache Hadoop and Mahout.

3. TOOLS & TECHNOLOGIES

From the literature review, it is well understood that most of the organizations exploit Mahout technology and Hadoop for data analysis. When implementing Big Data projects, selecting appropriate technologies is a part of processing the data sets. Almost everywhere we go online nowadays, Apache Hadoop plays a vital part in statistical analysis, ETL processing, and business intelligence. Popular website or web services like

Facebook, eBay, Yelp, and Etsyall uses Hadoop and Mahout to analyses their huge datasets. The organizations use Mahout to either generate data about the user's behavior or even for their own operations. This section details the working of Hadoop and Mahout in detail.

3.1 Hadoop

Google faced a problem of handling billions of searches and indexing millions of web pages. But, they could not find any distributed, large and scalable computing platforms for analyzing their own data. To address this problem, the Google team decided to create their own algorithm that allowed the large data calculations to be chopped up into smaller chunks and map into many computers. Google released their own Google File system paper [23] and MapReduce paper [24] in 2003 and 2004 respectively. Inspired by these papers, Doug Cutting created an open source project called, Hadoop. Doug Cutting is also the creator of Apache Lucene. Apache Lucene is a popularly used text search library [25]. Hadoop helps to store and process Big Data in a distributed cluster of computers using simple programming models. Hadoop is a part of Apache project which is sponsored by Apache Software Foundation. It is written in Java language for distributed storage and processing of mountains of data on computer clusters built from commodity hardware. The core concepts of Hadoop are Hadoop Distributed File System(HDFS) and Map Reduce algorithms. We will take a closer look at these two components.

3.1.1 HDFS

HDFS helps to store petabytes of data among multiple nodes and servers. HDFS [26] is a specially designed file system that allows multiple machines to store and process data from a single source. It consists of a Name Node and a Data Node which operates as a master-slave architecture. Name Node serves as a master component whereas Data Node serves as a slave component. Name Node comprises of metadata information of HDFS [26]. Metadata includes data like a number of blocks present in DataNode; a total number of times the data files being replicated; starting a cycle of the NameNode; the number of DataNodes which are under a particular NameNode; capacity of the NameNode; and information about space utilization. Data Node comprises information of all the processing data

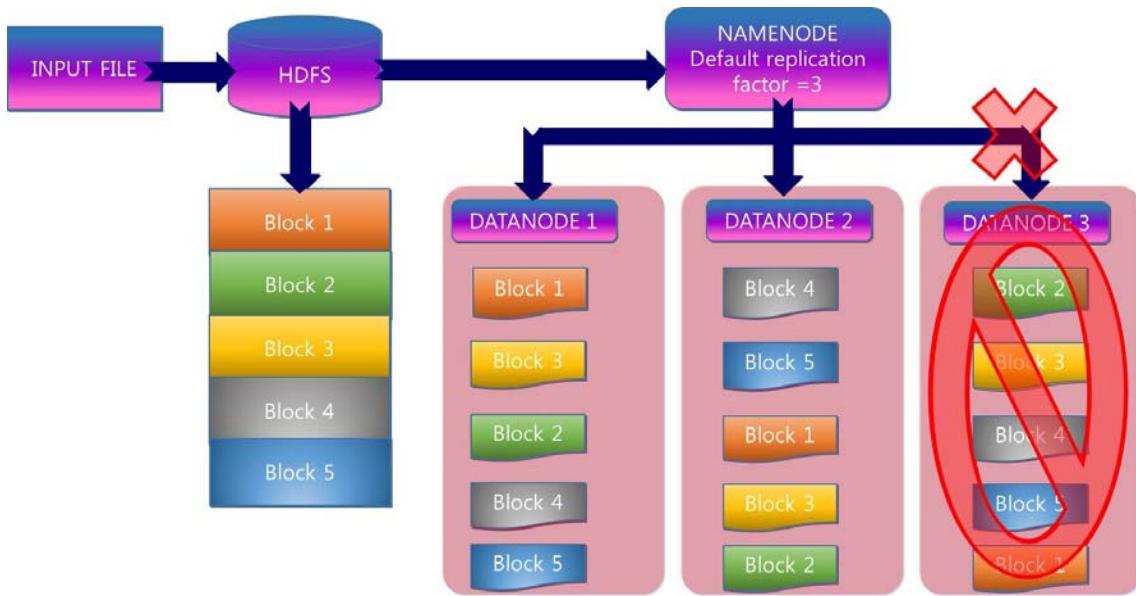


Figure 1: Replication process

which are stored in the Data Node and the machine it is being deployed. DataNode also deals with the actual storage of the file being processed and it serves READ and WRITE request for the client [27]. In the earlier versions of Hadoop, there was only one NameNode attached to multiple DataNodes which will result in a single point of failure. Later, the new versions of Hadoop provide multiple NameNodes where secondary NameNode can take over the entire process in case the primary NameNode fails. Secondary NameNode is responsible for conducting periodic checkpoints. HDFS also has self-healing architecture, where it replicates the same data across different nodes. So, it can process the data in a high availability environment. Let's say we have one NameNode and three DataNodes, and the data is transferred to the DataNodes from the client environment. HDFS has a replication factor which defines how many times a data block should be replicated in the clustered environment. For example, we have a file that is split into five data blocks across three DataNodes. The replication factor is set to three. If one of the nodes failed, the data from the failed node will get redistributed among the remaining active nodes. Hence, the other nodes complete the processing successfully. Figure 1 depicts the replication process in detail. In Hadoop 1.0 the NameNode uses Jobtracker and data node uses Tasktrackerto manage resources.

3.1.2 Map Reduce Map Reduce is a parallel programming model which helps to retrieve the data from Hadoop cluster and process all the data across commodity computer clusters. Processing data could mean anything like searching, counting the keywords, aggregating or enriching the data. Map Reduce is composed of several components. Some of them are **Job Tracker**: It is the master node which manages all the resources and jobs in the cluster. It also schedules the jobs and keeps track of the map reduce tasks across all the nodes. In case of failure, JobTracker reassigns the job to some other node. To sum up, JobTracker makes sure that the Big Data process successfully and the client get backs the data in a reliable condition. **Task Trackers**: They are agents that are deployed in each cluster to run the map reduce tasks which are assigned by the JobTracker. It sends a periodic heartbeat message informing that the node is running without any interruption. **Job History Server**: It is a component which tracks all the completed jobs and is deployed with JobTracker. The implementation of the program is two parts: Map and Reduce. The data is initially fed into the Mapper function to form intermediate key and value pairs. Once the mapping function is done, Reducer gets all the intermediate results to produce a final output.

3.2 Apache Mahout

Mahout is a machine learning library which is an open source from Apache. Apache Mahout aims to find out useful patterns from a large set of data, later predicting patterns from the new datasets as well. In this section, we shall discuss the fundamental concepts of Mahout clustering.

3.2.1 Distance Measure Distance measure technique is important implementation because the results of the same algorithm might differ depends upon the distance measuring technique. So, it is important to choose the distance measure based on the dataset. Distance measure affects the final quality of clustering. The data set to be clustered should be considered before selecting the distance measure. For clustering points which are well distributed in space, Euclidean distance measure suits well. If we need to cluster news snippets, cosine distance measure can be applied. For varying lengths of a document such as news articles, Tanimoto suits well. For complex clustering, like identifying friend root in the social network graph, weighted distance measure will apply. Different datasets have a different effect on the distance. If we knew the data set well enough, implementing our own distance measure will result in quality improvement. We used Euclidean, squared Euclidean, Tanimoto, Manhattan and cosine distance measure for clustering our datasets.

3.2.2 Representing Data Representation of input data as vectors help the algorithms to understand the data sets better and to calculate the similarity between the datasets. Datasets can be described as an ordered list of values, which is a vector. In 2-dimensions, vectors are represented as (5,8) first value for x dimension and the second one for y. Mahout can contain any number of dimensions when we deal with various kinds of input data. All the datasets contains features. Vectorization starts with assigning each feature to each dimension.

3.2.3 Vectorization Understanding the procedure of vectorization is essential to achieve accurate clusters. Apache takes vectors as input to produce good results. Vectors can be implemented as two classes which are explained below,

- Dense Vector: It implements the vector as an array of doubles because dense vector assumes that the size equals the number of features in the input datasets. It can handle zero values as well.
- Sparse Vector: Sparse vectors consists of two types named, Random and Sequential. The former

is implemented between a double and an integer. Random Sparse vector allocates non-zero features. The latter vector which is sequential is implemented in parallel arrays where one for integers and other for doubles.

However, the vector selection does not affect the final results but, will increase the calculation speed and decrease memory usage. The vectors are stored in the sequence file format, which can be read by Mahout algorithms. Sequence files are files which consist of key value pairs. Hadoop Map Reduce use input and output data sets in sequence file format. As Mahout runs on top of Hadoop, it is necessary to give the input in sequence file format. Among two kinds of dataset, one of the dataset we used is text format. We shall discuss the process of converting text into vector format. In real time scenarios, the text documents that are to be clustered will come in petabytes of sizes. And also all the documents will not be of uniform length. In the above subsection, we gave example for 2-dimension vectors. But, in the real case, the dimension of the text document vectors will be really large. Since the count of words in a document of size say terabyte and petabyte will be incredibly big, the vectors will have infinite dimensions. One important factor to be considered is the frequency of words in a given file. A number of occurrences of a word are referred as a value of the vector. This is defined as Term Frequency (TF) weighting. In Mahout, the value of the vector is technically referred as weight. In text document clustering, the similarity between any two documents is calculated based on distance measure. TF does not know about the common words like “the”, “is”, “a” and it calculates the weight by including all the words in the document. So the result will be disagreeable, as the distance value will be influenced by the weight of the frequent words. Such frequently occurring words are referred to as stop words. This stop-words problem can be overcome by Term Frequency- Inverse Document Frequency (TF-IDF) which is discussed in the following section.

3.2.4. TF-IDF Term frequency-inverse document frequency (TF-IDF) weight is a statistical measure used to detect the importance of a particular term in the document. Intuitively, a word is given high score if it occurs frequently in a document. But, the word is given low score if it occurs frequently in many documents. Thus, we can conclude that the word is not unique identifier. Words like “the”, “for”, “a”, “what” should be

scaled down since they are most common words which occur across multiple documents, whereas words which occurs frequently in single document should be scaled up. In TF-IDF, Inverse document frequency (IDF) is used to improve the weighting. DF is the number of documents a word occurs in and N is the total number of documents present. The log value is determined to reduce the problem of final term weight as the value wont be ideal. The formula for IDF for a particular word, w_i is mentioned below.

$$IDF_i = \log(N/DF_i)$$

Now, to calculate the TF-IDF weight for a word w_i is,

$$W_i = T F_i * \log(N/DF_i)$$

This above calculated weight for a word I will be assigned as a dimension. By this, stop words will have small weight whereas the unique words will get larger weights, resulting in best cluster formation. Vector space model assumes that all the words in the document are independent to each other. Still, word dependencies are another threat for weight calculation. Some words like "coca cola" has a high probability of occurring together which is called word dependencies. In the next subsection, we shall discuss how to overcome the word dependencies.

3.2.5 Normalization n-gram in general is defined as group of words that occur in a sequence. Combination of single word is defined as "unigram" and similarly combination of two words is called "bigram". More than two words combinations are referred as "trigram", "4-gram" and so on. Mahout has a technique to find out the highest probability of combined words like coco cola, Martin Luther. Vectors can be created to point at bigrams instead of single words. From a single sentence, bigrams can be created by picking all the possible combination of two words. Some may give meaningful units, while most of the bigrams represent worst meaning [28]. If bigrams are created for all the documents, eventually we will end up with lots of meaningless bigrams with big TF-IDF weight. To outdo this problem, Mahout conduct log-likelihood test. This test can conclude whether the bigrams forms notable meaning or not. It selects the bigrams with most appropriate meaning and eliminates the remaining. With the help of TF-ID weight and n-gram collocation, text documents are converted into

input vectors. Thus resulting in high quality of clusters.

The text document is converted into sequence format for the sake of HDFS by using the Mahout launcher script. Later, the sequence format of input data is converted into vector for better cluster formation. Once the vector is formed from the input data, the following output are written in HDFS as a result of vectorization process.

- Tokenized documents:** The documents are segregated as individual words, using Standard Analyzer and stored in this file.

- Word Count:** The n-gram is generated by iterating through the tokenized documents and the important meaningful words are stored in this folder.

- TF vectors:** Using term frequency , TF vectors are created from the tokenized documents.

- DF count:** The frequency of the words across multiple documents is calculated for TF-IDF weight.

- TF-IDF vectors:** The vectorizer uses TF-IDF weight, so after the generation of TF vectors, the weight for TD-IDF is calculated.

- Dictionary file:** The final folder "dictionary.file-0" consists of mapping between a word and the integer. This file is used as input for the various algorithms in Mahout. This file contains all the meaningful bigrams, trigrams as well.

Normalization is a method for cleaning up the edge cases. Data with ab-normal characteristics change the direction of the resulting clusters disproportionately. In general, it is common for any two documents which are not all similar but they pop up claiming they are similar. This is due to the size of the document. When a larger document is being compared with a smaller document, it will results as they are similar because of the large number of non-zero dimensions. The error should be negated as the incoming datasets are not of same length and size. Normalization decreases the magnitude of the large vectors and increases the magnitude of the small vectors. The quality of the final clusters are increased to an extent by normalization. Further increase in clusters quality depends upon the selection of distance measure and suitable algorithms for the input datasets. Distance measures were already discussed in the previous section. The working of the clustering algorithms are exhaustively explained in the following section.

4. CLUSTERING ALGORITHMS

Clustering in Mahout refers to organizing the unknown elements based on their similarity. In this section, we present a brief overview of some important clustering algorithms in Mahout. The general working of each algorithm and the way it works in Mahout as well are discussed below.

4.1 K-means Clustering

K-means is the traditional simple algorithm which ages 50 years old. As the name illustrates, k is the important parameter which is used to set the number of output clusters [29]. The quality of the cluster is dependent on the value of k. In our paper, one of the dataset we clustered is news dataset. The expected out-put should be grouped into various news categories say politics, health, sports and the like. The first step is to assign the centroids among the data points. Mahout chooses the initial centroid randomly by random seed generator. The second step is to assign the data points to the nearest centroids to form the initial clusters and calculate the mean. The above steps are repeated again and again until the data points are assigned to the same cluster in successive rounds. Figure 2 shows the working of the K-means algorithm as a flow chart. Mahout computes the K-means algorithms as Expectation Maximization (EM) algorithm [30]. Based on the initial centroids, the expected points are assigned to the cluster. Later the cluster center is improved by varying distance measure until it reaches the convergence point. Convergence point is that point where the centroids don't move further after the computing. Then no further iterations are required and the final clusters are displayed. In Mahout, convergence threshold, maximum iterations, distance measure are used as necessary parameters to execute the K-means algorithm. K-means will read from the dictionary file, which is created as a result of vectorization and form the final clusters.

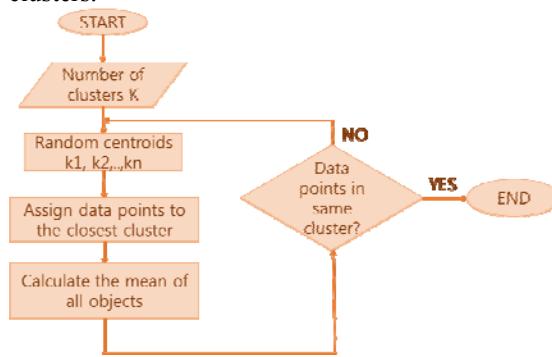


Figure 2: Work flow of K-means algorithm

K-means algorithm can be executed both by java code and MapReduce. Java implementation is feasible only up to certain limit of dataset. When dealing with large amount of data, MapReduce gives best result in clustering. As Hadoop and MapReduce splits the input into several chunks for parallel processing, Mahout framework takes care of the spilt chunks of the vectors. After the formation of vectors, the algorithm command is executed with all the necessary input parameters. After the implementation of K-means, the output directory contains all the clusters which are formed from first iteration until the final iteration. Clustered Points directory consists of the mapping information for documents and cluster. Figure 3 shows the final output directory of cluster formation in HDFS. The final clusters formed in the directory is in vector format which is not understandable by humans.

hdfs://namenode/output							
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rwxr--r-	hduser	supergroup	194 B	5/11/2016, 5:48:19 PM	1	128 MB	_policy
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:48:43 PM	0	0 B	clustersPoints
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:48:45 PM	0	0 B	clusters0
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:14 PM	0	0 B	clusters1
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:25 PM	0	0 B	clusters2
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:31 PM	0	0 B	clusters3
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:37 PM	0	0 B	clusters4
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:42 PM	0	0 B	clusters5
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:44 PM	0	0 B	clusters6
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:46 PM	0	0 B	clusters7
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:48 PM	0	0 B	clusters8
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:50 PM	0	0 B	clusters9
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:52 PM	0	0 B	clusters10
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:54 PM	0	0 B	clusters11
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:56 PM	0	0 B	clusters12
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:58 PM	0	0 B	clusters13
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:59 PM	0	0 B	clusters14
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:59 PM	0	0 B	clusters15
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:59 PM	0	0 B	clusters16
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:59 PM	0	0 B	clusters17
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:59 PM	0	0 B	clusters18
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:59 PM	0	0 B	clusters19
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:59 PM	0	0 B	clusters20
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:59 PM	0	0 B	clusters20_final
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:59 PM	0	0 B	clusters3
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:59 PM	0	0 B	clusters4
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:59 PM	0	0 B	clusters5
-rwxr--r-	hduser	supergroup	0 B	5/11/2016, 5:49:59 PM	0	0 B	clusters6

Figure 3: Clusters output directory

Mahout provides a utility called ClusterDumper which can convert vector format into words and dump them in the local directory.

4.2. Canopy Clustering

Canopy is not only fast clustering process but also accurate in grouping the data points based on some sort of similarity. It mainly uses two distance measures named T1 and T2 where $T1 > T2$. The first step is to select any point randomly from the input datasets to form the canopy center. From that random point the distance to all other data points is calculated. By keeping all the points which falls into the distance value $T1$ it forms an initial cluster. Remove all the data points that fall within the distance value $T2$. These points can neither be a new centroid nor be capable of forming new canopies. The above steps are

repeated again and again till there are no more points to form a new cluster.

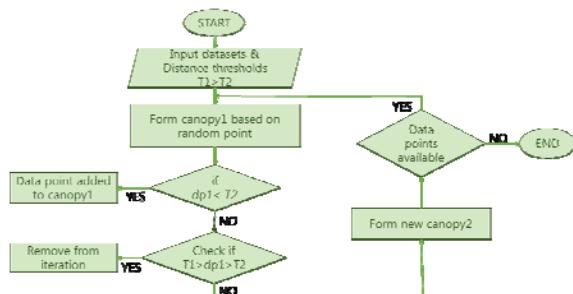


Figure 4: Flow chart of Canopy algorithm

The advantage of canopy clustering is its quick formation of clustering, but the quality of the final clusters cannot be guaranteed. The one disadvantage in K-means is the random initial centroid formation. If the initial centroid selection is not good, then again the quality of the final clusters might be bad. In order to overcome both the disadvantages, Mahout provides a way to combine canopy and k-means algorithm. Canopy algorithm is used to form the centroids instead of random seed generator.

The dataset about news articles contain may overlapping topics which cannot be clustered under a single topic. Let's say, a news article which talks about celebrity's death might be clustered in both health and movies. In such case, the algorithm should be able to replicate the news article in both the categories. In such cases, the fuzzy K-means clustering algorithm serves better which is discussed in the following section.

4.3. Fuzzy K-means Clustering

Fuzzy k-means otherwise known as fuzzy c-means clustering algorithm. This algorithm tries to create overlapping clusters between the datasets. Fuzzy k-means working is based on soft cluster where one data point can belong to more than one cluster.

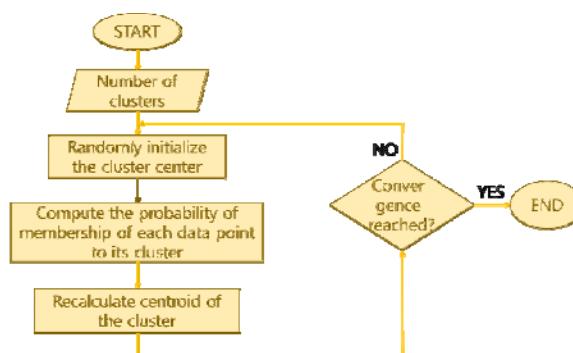


Figure 5: Working flow of fuzzy k-means algorithm

In Mahout, the distance measure, convergence threshold, number of iterations and fuzziness factor are given as input parameters. The value of the fuzziness factor should be greater than 1. As the fuzziness factor becomes greater than 1 the datasets might have more overlap. Figure 5 gives overall working of fuzzy k-means algorithm. The algorithm runs in MapReduce mode involving the following three steps.

- Mapper: Reads the input dataset and calculate the membership probability of the data point to its cluster.

- Combiner: Receives the output from mapper and calculates the partial sums of probability of each vector.

- Reducer: Calculates the total sum of the membership probability and creates a new centroid for the cluster.

If the data is uncertain, fuzzy k-means gives better cluster quality than K-means. There is one disadvantage that either of the above three algorithms cannot handle. Standard k-means is well suitable for exclusive clustering and fuzzy k-means is for overlapping clustering. But if the datasets are not in the normal distribution, then applying the above algorithms might not generate a best fit for the data points. For an instance, to form a cluster based on specific topic which is debatable. A word can be included in two topics because it has same meaning, but not both. To prevail over this issue, topic modeling algorithms are implemented which are discussed further in the upcoming sections.

4.4. LDA

Latent Dirichlet Allocation (LDA) clustering algorithm learns the pattern from the input data based on the similarity of the context, it forms the final clusters. In the dense context, LDA assumes that the datasets has t number of topics and the entire input datasets are revolving around the topic. Single input document is considered to have mixed probabilities of each t number of topics. In topic modeling, the algorithm takes word vector as input instead of document vector as it deals with each word of the document to find out the probability of the topics. In the word vector, IDs are features and the total number of documents they occur in will be considered as weights. Using word vector, the cluster of the words are figured out by executing the algorithm and the topics are discovered. After the discovery of the topics, the probabilities of each word occurring in that topic is calculated. This is the general working of topic modeling.

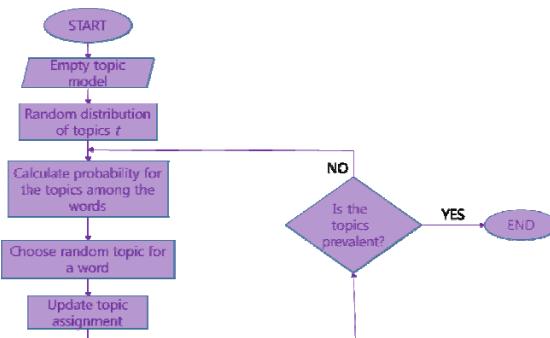


Figure 6: Flow model of LDA algorithm

LDA works one step ahead of topic modeling. Even if any words which have the same meaning but do not occur together, LDA scrutinize across the documents, to figure out the similar words and the content of the document. After examining the words which have the same concept, the words are considered as a single topic. The following figure 7 shows the required input to be given to run LDA algorithm. The matrix results represent the most frequently occurring words in the documents. This is later used for building an LDA model. The model based algorithm, LDA creates good quality of clusters when the datasets are static. The prediction of pattern in the data and the creation of cluster quality is high when we deal with a large amount of stored data set. But, when dealing with real-time datasets, finding a pattern from the dataset might be difficult if the data keeps on accumulating in the directory.

The screenshot shows a Hadoop interface with a green header bar containing 'Hadoop', 'Overview', 'Datanodes', 'Snapshot', 'Status Progress', and 'Utilities'. Below the header is a 'Browse Directory' section for 'Reuters-pta'. It displays two files: 'documents' (1.27 kB, modified 8/23/2016, 6:04:59 PM) and 'matrix' (1.1 MB, modified 8/23/2016, 6:04:59 PM). At the bottom, a terminal window shows the command 'Hadoop, 2015.'

Figure 7: Output of LDA

To overcome this problem, streaming k-means is used which is detailed in the following section.

4.5. Streaming K-means Clustering

Streaming K-means clustering algorithm is most suited for dynamic kind of data [31]. It is mainly used for online clustering, where the input data points are received every second. In our paper, we do not deal with real time data.

5. Experiments

5.1. Datasets

In order to compare the results of the algorithms, two different types of datasets are used, text and xml. We downloaded bag of words dataset for the former and Multivariate for the latter. The size of the text and xml dataset are 7.27 GB combining both test and training data uncompressed.

5.2. Setup

We had 8 node Hadoop clusters which is based on commodity hardware. Each node has the following same configurations as follows: 8GB RAM; 4 integrated Gigabyte Ethernet 1000BASE-T ports(RJ-45); 0.5 TB hard drive; 1CPU Inter Core i4; 30 MB L3 cache with Ubuntu version 14.04, Hadoop v2.7and Mahout v0.9. The evaluation is detailed corresponding to distance measure, convergence and clustering speed.

5.3. Use cases

The following are the different scenarios we considered for doing the Mahout case study. (i) Text dataset 7 GB tested with different distance measures. (ii) XML dataset 6 GB tested with different distance measure. (iii) Text dataset with K-means; Single node; experimented by varying the convergence threshold.(iv) XML dataset with Fuzzy K-means; Single node; experimented by varying the fuzziness factor. (v) Text dataset clustering with LDA; pseudo distributed with 1 server and 8 nodes.

6. RESULTS & INTERPRETATION

In this section, the analysis results of the clustering algorithms are discussed exhaustively. We considered four clustering algorithms for the interpretation: K-means, K-means with Canopy, Fuzzy K-means and Latent Dirichlet Allocation (LDA). The above mentioned datasets were loaded into HDFS. The block size of HDFS was 128 MB. The replication factor of the dataset is set to default value which is 3. The input parameters such as initial cluster, total number of clusters, distance measure, distance thresholds make a huge difference in the resulting quality of the clusters and the accuracy.

Table 1: Clustering parameters

PARAMETERS	Text Values	XML values	Fixed clusters
K-means	20	20	Yes
Canopy Clustering T1	2050	2000	No
Canopy Clustering T2	1050	1500	No

Fuzziness factor m	-m 1.05, 2, 2.5, 3.0	2	Yes
-----------------------	-------------------------	---	-----

In K-means, selecting the number of clusters (k) depends upon the input data and the expected output. In our text dataset, the value of the parameter k should be 10 to 20 if we want to cluster only general topics say sports, health, politics etc. On the other hand, if the resulting category needed to be finely clustered, then the input number of clusters should be increased to 35 to 40. If the volume of data is too large to be stored in the main memory available, the K-Means algorithm not suitable, as it's batch processing mechanism iterates over all the data points. Also, the K-Means algorithm is sensitive to the noise and outliers in data. Moreover, its random initialization step causes problems when it comes to computation time and accuracy. Canopy clustering has been applied for both text and XML datasets. We used canopy clustering as a pre-clustering algorithm. The result of the canopy clustering was used to estimate the values of centroids and number of clusters. Later, the initial set of clusters has been used as centroids in K-means algorithm. Canopy integrated with K-means gives quick and accurate results. From the results shown in the graph8, it is clear that canopy with k-means clustering gives quick results for XML dataset than text dataset. For large dataset which has high dimensions, canopy clustering does not work well. This is because high dimensional data points can be involved in many canopies, thus resulting in poor quality of cluster formation.

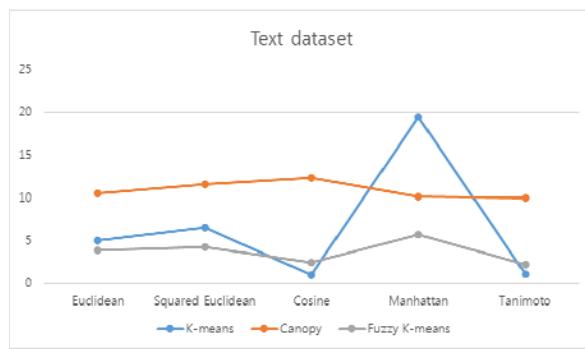


Figure 8: Clustering algorithms results of text dataset

Figure 8 depicts the processing speed of clustering algorithm by varying the distance measure. The X-axis shows the distance measure while the Y-axis indicates the time taken for the processing of the clusters in minutes. From the graph it is clear that Manhattan distance measure stands out extraordinarily huge from all other

distance measures apart from the convergence value. The maximum iterations for all the algorithms have been kept constant as 20. For a small size dataset like 2GB, Manhattan distance measure seems to be taking very long time. Surprisingly, for high dimensionality vectors Manhattan with K-means algorithm gives better quality cluster. It is clear that fuzzy K-means algorithm suits well for text dataset than XML dataset. Text dataset has to be soft clustered rather than hard clustering. The nature of the news datasets that they cannot be categorized in one single category. The news which talks about health issues of a sportsman should be clustered in both health category and sport category. The Euclidean distance measure is not fit for the text documents of varying size. This is because, the distance measure fails to consider the length of the documents. Hence, Fuzzy K-means is best suited for our text dataset. Except Manhattan distance measure, all others form the final clusters more or less in a same time period. The maximum iteration parameter for all the distance measures and the clustering algorithms is set constant values 20.

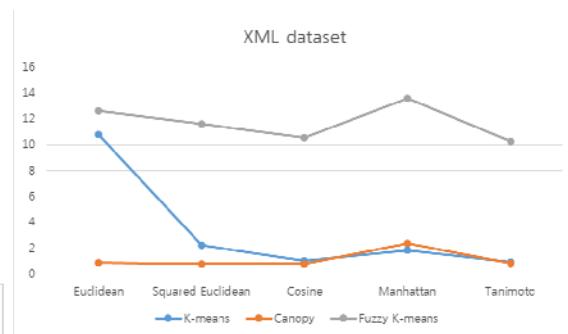


Figure 9: Clustering results of XML dataset

The figure 9 shows the results of the various clustering algorithm for XML dataset executed using different distance measure. The X-axis indicates various distance measure while the Y-axis shows the time taken in minutes for the formation of the clusters. Overall from the graph it is shown that canopy clustering algorithm works well for XML dataset. As Fuzzy K-means algorithm has overlapping nature, it showed better and quick results for text dataset. We expected the same kind of results for XML dataset. But we observed that the cluster formation takes much longer time even though we minimized the dataset to 2 GB. The cluster quality of fuzzy K-means was not good as most of the data points have been assigned to almost all the clusters. Figure 10 shows

the clustering results of the text dataset by varying the convergence threshold.

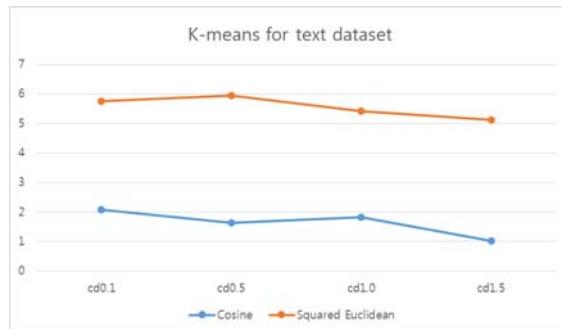


Figure 10: Clustering results of text dataset by varying the convergence threshold.

Hence, the centroids were not properly representing the clusters. In this case, even if the centroids are randomized, K-means clustering algorithm produced better clusters for XML dataset. With canopy as centers, K-means converge even faster irrespective to distance measure.

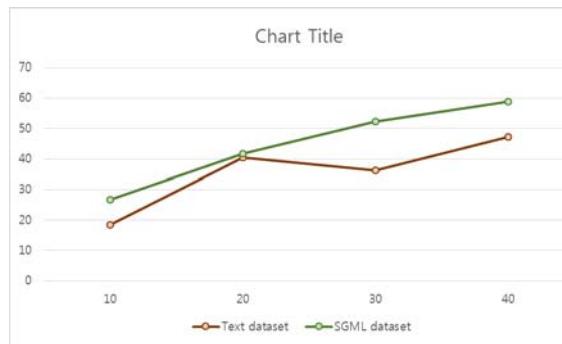


Figure 11: LDA model results of both text and sgml dataset

The results of the figure 11 indicates the k-means clustering results by varying the convergence threshold for text dataset. The X-axis indicates the values of convergence threshold while the Y-axis shows the execution time in minutes. It is clear that using cosine distance measure for clustering makes the dataset converge much faster when compared to squared Euclidean distance measure. When the distance measure we used is cosine, we observe that as the convergence value increases, the clustering time gets decreased. But, In case of squared Euclidean distance measure, none of the clusters reached convergence. All the clusters formed 10 clusters because it reached maximum iterations. It is known from the results that the squared Euclidean distance measure need

more convergence for the resulting clusters. In the case of cosine distance, the number of resulting clusters also varies based on the convergence factor. Anyways, K-means clustering initialize its centroid using random seed generator. So, the results may vary when we use canopy centroids. From the table 2, it is visible that as the fuzziness factor gets increasing, there is slight decrement in the execution time. Though in our XML dataset, none of the clusters reached convergence. All the clusters reached maximum iterations and got stopped. From this results it can be decided that fuzzy k-means algorithm is not best suited for XML dataset.

Table 2: Clustering results for XML by varying the fuzziness value

Parameters	Fuzzy factor	Time taken (mins)
Cd 1.0, dm Squared Euclidean	-m 1.5	5.8829
	-m 2.0	5.6485
	-m 2.5	5.5325
	-m 3.0	5.3348

The results of the figure 11 shows that LDA gives quick results in case of text analysis. The X-axis indicates the maximum iteration whereas Y-axis indicates time in minutes. The input for the LDA analysis is raw text, and LDA models topics out of the raw input data. The table 3 lists top three topics among all the results and the corpus with highest probability values is displayed here. From the group of words, the topics they might be allocated is quickly understandable.

Table 3. Results of Topic model in LDA using 20 newsgroup dataset

Topic 1	Probability	Topic 2	Probability	Topic 3	Probability
Food	0.018	Space	0.023	Leaders	0.021
Treatment	0.012	Earth	0.012	Government	0.011
Cancer	0.0009	Launch	0.009	People	0.009
Medicine	0.006	Nasa	0.005	Organization	0.005
Health	0.006	Shuttle	0.004	Law	0.004
Disease	0.005	Research	0.003	Policy	0.002

Topic 1, 2 and 3 talks about health news, space exploration and politics respectively. LDA is successful because of its ability to produce interpretable topics among large amount of documents. Still, parallelization in LDA is not straightforward. From the graph 11, it is evident that none of the lda model attains convergence. The topics keep on modeling till the maximum iteration which is 40 in our case. Using EM

algorithm the convergence can be achieved in short number of iterations.

7. CONCLUSION

Mahout and Hadoop are promising platform for large data analysis due to open source and scalable characteristics. In-built libraries of Mahout and parallel processing feature of Hadoop provide better quality of resulting clusters. From the paper, it is clear that not all the algorithms are suited for all kind of datasets. There are many factors such as distance measure, convergence threshold, iterations, number of nodes to be considered before feeding the input datasets. We have experimented with text dataset and XML dataset. Although the selection of algorithm entirely depends upon the types of dataset, we still can choose the best suited distance measure from the above results.

Future work includes applying classification and recommendation algorithms for different kinds of datasets like CSV (Comma Separated Values).

8. ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2016R1A6A3A11932892)

REFERENCES:

- [1] X. Zhu, B. Song, Y. Ni, Y. Ren, R. Li, Business Trends in the Digital Era: Evolution of Theories and Applications, Springer, 2016.
- [2] J. A. Hartigan, J. Hartigan, Clustering algorithms, Vol. 209, Wiley New York, 1975.
- [3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers, Big data: The next frontier for innovation, competition, and productivity.
- [4] V. M. Rohokale, N. R. Prasad, R. Prasad, A cooperative internet of things(iot) for rural healthcare monitoring and control, in: Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE), 2011 2nd International Conference on, IEEE, 2011, pp. 1–6.
- [5] S. Loi, N. Sirtaine, F. Piette, R. Salgado, G. Viale, F. Van Eenoo, G. Rouas, P. Francis, J. P. Crown, E. Hitre, et al., Prognostic and predictive value of tumor-infiltrating lymphocytes in a phase III randomized adjuvant breast cancer trial in node-positive breast cancer comparing the addition of docetaxel to doxorubicin with doxorubicin-based chemotherapy: Big 02-98, Journal of Clinical Oncology 31 (7) (2013) 860–867.30
- [6] I. S. Kweon, T. Kanade, High resolution terrain map from multiple sensor data, in: Intelligent Robots and Systems' 90. Towards a New Frontier of Applications', Proceedings. IROS'90. IEEE International Workshop on, IEEE, 1990, pp. 127–134.
- [7] K. Coussette, D. Van den Poel, Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques, Expert systems with applications 34 (1) (2008) 313–327.
- [8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (Oct)(2011) 2825–2830.
- [9] S. Owen, S. Owen, Mahout in action.
- [10] W. Romsaiyud, Big data topic modeling with mahout for managing business analysis services, in: System Integration (SII), 2014 IEEE/SICE International Symposium on, IEEE, 2014, pp. 514–519.
- [11] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation, Journal of machine Learning research 3 (Jan) (2003) 993–1022.
- [12] R. Irudeen, S. Samaraweera, Big data solution for sri lankan development: A case study from travel and tourism, in: Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on, IEEE, 2013, pp. 207–216.
- [13] A. D. Barrachina, A. O'Driscoll, A big data methodology for categorising technical support requests using hadoop and mahout, Journal of Big Data 1 (1) (2014) 1.
- [14] F. Xhafa, D. Ramirez, D. Garcia, S. Caballé, Performance evaluation of data mining frameworks in hadoop cluster using virtual campus log files, 31 in: Intelligent Networking and Collaborative Systems (INCOS), 2015 International Conference on, IEEE, 2015, pp. 217–222.
- [15] C. Rong, et al., Using mahout for clustering wikipedia's latest articles: a comparison between k-means and fuzzy c-means in the cloud, in: Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on, IEEE, 2011, pp. 565–569.
- [16] L. Sahu, B. R. Mohan, An improved k-means algorithm using modified co-sine distance measure for document clustering using mahout with hadoop, in: 2014 9th International Conference on

- Industrial and Information Systems (ICIIS), IEEE, 2014, pp. 1–5.
- [17] S. Rallapalli, R. Gondkar, G. V. M. Rao, Cloud based k-means clustering running as a mapreduce job for big data healthcare analytics using apache mahout, in: Information Systems Design and Intelligent Applications, Springer, 2016, pp. 127–135.
- [18] C. I. Muntean, G. A. Morar, D. Moldovan, Exploring the meaning behind twitter hashtags through clustering, in: Business Information Systems Workshops, Springer, 2012, pp. 231–242.
- [19] E. Jain, S. Jain, Categorizing twitter users on the basis of their interests using hadoop/mahout platform, in: 2014 9th International Conference on Industrial and Information Systems (ICIIS), IEEE, 2014, pp. 1–5.
- [20] T.-S. Moh, S. Bhagvat, Clustering of technology tweets and the impact of stop words on clusters, in: Proceedings of the 50th Annual Southeast Regional Conference, ACM, 2012, pp. 226–231.
- [21] C.-L. Yang, M. R. Nurtam, Data clustering on taiwan crop sales under hadoop platform, in: Proceedings of the Institute of Industrial Engineers Asian Conference 2013, Springer, 2013, pp. 827–835.
- [22] V. R. Eluri, M. Ramesh, A. S. M. Al-Jabri, M. Jane, A comparative study of various clustering techniques on big data sets using apache mahout, in: 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC), IEEE, 2016, pp. 1–4.
- [23] S. Ghemawat, H. Gobioff, S.-T. Leung, The google file system, in: ACM SIGOPS operating systems review, Vol. 37, ACM, 2003, pp. 29–43.
- [24] J. Dean, S. Ghemawat, Mapreduce: simplified data processing on large clusters, Communications of the ACM 51 (1) (2008) 107–113.
- [25] T. White, Hadoop: The definitive guide, "O'Reilly Media, Inc.", 2012.
- [26] D. Borthakur, et al., Hdfs architecture guide, Hadoop Apache Project 53.
- [27] P. Zikopoulos, C. Eaton, et al., Understanding big data: Analytics for enterprise class hadoop and streaming data, McGraw-Hill Osborne Media, 2011.
- [28] N. K. Visalakshi, K. Thangavel, Impact of normalization in distributed k-means clustering, International Journal of Soft Computing 4 (4) (2009) 168–172.
- [29] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, et al., Constrained k-means clustering with background knowledge, in: ICML, Vol. 1, 2001, pp. 577–584.
- [30] F. Dellaert, The expectation maximization algorithm, 2002, URL <http://www.cc.gatech.edu/dellaert/em-paper.pdf>.
- [31] K. Venkatalakshmi, S. M. Shalinie, Classification of multispectral images using support vector machines based on pso and k-means clustering, in: Proceedings of 2005 International Conference on Intelligent Sensing and Information Processing, 2005., IEEE, 2005, pp. 127–133

Implementation challenges in Campus Network security

Zaka Ullah

Department of Computer Science
Lahore Garrison University, Lahore, Pakistan
Email: zakaullah@lgu.edu.pk

Muhammad Zulkifl Hasan

Department of Computer Science
Lahore Garrison University, Lahore, Pakistan
Email: zulkifl.hasan@lgu.edu.pk

Abstract—Recently, due to economical escalation & with the need of advancements in national strategy based on scientific technologies, maximum institutions have been adopting their own campus network. That is why setting up campus network is being realized as important part of school education information. The utilization of campus network shares opportunity for teaching, scientific research and management to work together by using resources and exchange maximum information in minimum time. Hence, campus network security positively influences school teaching activities.

Index Terms—SSL, PKI, IDS, IPS, VPN, URL, DNS, DMZ, LAN, IT.

I. INTRODUCTION

With the expansion of society, organize data is utilized by individuals to an ever-increasing extent and its significance is getting to be prominent, the system has turned into the most basic thing of a nation's political, financial, and military assets. Nowadays security is compromised due to various applications and can cause data spillage, illegal data entrance, and unlawful utilization of system assets, data altering and counterfeit. The security may include identification, authentication, authorization, and surveillance camera to protect integrity, availability, accountability, and authenticity of computer hardware or network equipment [Ali et al.(2015)Ali, Hossain, and Parvez]. As the pervasiveness of multi-PC arrange security dangers, avert "programmers" is feeble, corporate and government sites have been "assaulted" increasingly often, these have caused tremendous monetary misfortunes. Subsequently, the system data framework security and avoidance and its Secrecy appears to be progressively essential. With the quick extension of the grounds arrange availability, the system applications have expanded quickly, in the meantime, the grounds organize data security has caused more consideration today. A hierarchical architecture of campus network is configured with different types of traffic loads and security issues for ensuring the quality of service. [Ali et al.(2015)Ali, Hossain, and Parvez] In recent organize and application frameworks, there are no safety efforts which have been taken care off. Meanwhile security vulnerabilities in the host working framework and application framework are additionally with no handling, there are numerous issues inside framework administration, all of these shaped a genuine security issue, therefore truly debilitating the security of the grounds arrange. Currently the framework

and the host was observed to endeavour to be attacked by others, a substantial number of security vulnerabilities exists in the framework, and there are numerous security vulnerabilities which are hard to keep away from and destroy; Also, virus transmitted through the system seriously influenced the ordinarily running of the grounds organize. Therefore, network architecture and its security are vital issues for any university. In this work, a network infrastructure is proposed on the basis of the practical and experimental requirements. The proposed network infrastructure is realizable with adaptable infrastructure [Das and Behera(2017)].

II. CHALLENGES IN DESIGNING CAMPUS NETWORK

School and University are generally furnished with a PC room; there are a few PCs which have open access to the grounds PC organize in this room, understudies also, personnel are normally accessible to utilize these registers to access to web to get data and learn on the web. The goal is to minimize cost based on these elements while delivering service that does not compromise established availability requirements. There may be two primary concerns: availability and cost. These issues are essentially at odds. Any increase in availability must generally be reflected as an increase in cost. [Das and Behera(2017)] However, the absence of bound together administration programming and framework for observing and logging, these PC rooms can't be basic in the administration state. Most rooms have genuine imperfections in enlistment and administration framework, so the web client's character cannot be perceived. It also must scale, offer operational simplicity and flexibility, accommodate new computing trends without much alterations in the original structure and design.

[K C Ramakrishnegowda(2009)] Its very convenient for us to use functions provided by the campus network, but it also has become a quick way to transfer the virus. The virus attack can invade the confidentiality of user and can results in information leakage. Network virus can deteriorate network resources that can decrease network performance rapidly. Hence network security is most vital component in information security, because it is responsible for securing all information passed through networked computers.[MADJE(2017)]

In the grounds organize, assaults, interruption the machines, robbery of another record, the unlawful utilization of the

system, illicit access to unapproved documents, badgering via mail and different episodes frequently happen, et cetera, demonstrating the clients' security awareness in the grounds organize are exceptionally unremarkable. Presently the internet is flooded with the hacking tools; hackers use network protocols, server and operating system security vulnerabilities and management oversight to illegally access to network resources, deletion of data, damage the system, these attacks caused to the adverse effects of the campus network and the damage to the reputation of the school.[Gursimrat Singh(2013)] A Most difficult issue is the deficiency of assets for the organize development in school or college, restricted assets are fundamentally put resources into arrange gear, efficient contribution for development and administration of system security has not been considered genuinely. In light of absence of the mindfulness in significant colleges, administration organizations are not perfect, administration framework is flawed, administration innovation is not propelled; these elements influence the college's system administration to focus can't take any measures and preventive measures for data security. In the interim, Countries don't have all around created and thorough system security framework, there is no strict execution of gauges for the grounds organize security administration; this is an imperative purpose behind the multiplication of system security 1 I is the normal sorts of system security dangers.

By the investigation over, the grounds arrange security issues is for the most part in the accompanying areas: Password exposure can bring about information spillage. An assortment of database frameworks is running on-line in the grounds organize, for example, showing administration framework, understudy accomplishment administration framework, grounds card administration framework, test bank et cetera. The overall topology is based on tree topology which is conveniently serving the network. The network is based on client-server architecture.[Begum and Nandyal()] The client individual offense or carelessness of wellbeing measures can prompt these database secret key be lost, the information might be illicitly evacuated or repeated, bringing about data divulgence, in genuine condition, may bring about genuine unlawful removal of information. Subsequently, setting secret key is likewise imperative.

Grounds systems can associate with the web with switches; obviously, web clients can appreciate the accommodation of quick and boundless assets of this stage, yet in addition need to face to the danger of an assault. The overall topology is based on tree topology which is conveniently serving the network. The network is based on client-server architecture. There is significant security chance inside the grounds organize, interior clients are generally see more about the arrange structures and utilizations of models than the outside clients, subsequently the interior security dangers are the principle dangers. As discussed earlier, hacking devices overwhelmed in the Internet, programmers utilize arrange conventions, server and working framework security vulnerabilities and administration

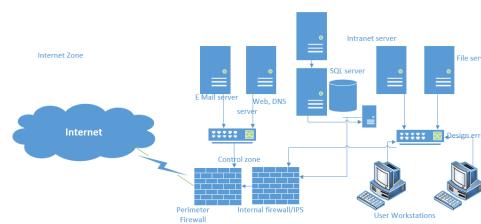


Fig. 1. Simulation results for the network.

oversight to illicitly access to organize assets, erasure of information, harm the framework, these assaults caused to the unfavourable impacts of the grounds system and the harm to the notoriety of the school.

There is a great deal of security vulnerabilities inside authentic working framework, these security vulnerabilities posture numerous genuine dangers, for example, the data security, utilizing of the framework, arrange operations et cetera. Security policy weaknesses can create unforeseen security threats. [Alabady(2009)] Attributable to the absence of attention to copyright, robbery programming, movies and TV assets is utilized as a part of general in the grounds arrange, while the spread of the product takes up a considerable measure of system data transmission; on the other hand it brings a specific system security dangers. Test case, Microsoft's 8.1 working framework organization has few extent of PC applications which are step by step extended in the grounds, get to focus to the grounds arrange have expanded to a great extent, however a large portion of these hubs is the new and a portion of the defensive measures have not been received, this may cause the virus flooding whenever, data misfortune, information debasement, arrange assaults, System crumble and other genuine results. Constraint in refreshing the product; if the clients introduce Microsoft's 8.1 working framework with the pilfered one, the PC framework will leave countless vulnerabilities. Then again, downloading from the Web programming which may has shrouded Trojans virus, indirect accesses and different malevolent code, in this way, numerous frameworks are intrusive and utilized by the aggressor.

A portion of the aggressors and merchants send spam and spread hurtful data utilizing unmanaged grounds server as a travel station, this has conveyed a genuine impact to the organize and expedited incredible impact the school's notoriety. In rundown, the security circumstance of the grounds organize is intense, keeping in mind the end goal to understand these security dangers and vulnerabilities, as indicated by the basic highlights of the grounds system and security issues which the grounds arrange confronted, security arrangements ought to be taken to the grounds system and it ought to be executed as snappy as could reasonably be expected.

III. 3 CAMPUS NETWORK SECURITY SOLUTIONS

The campus network safety provides shield for IT system resources against external threats like intruders and malicious code. The network security system includes firewalls, intrusion detection and prevention system (IPS/IDS), VPN protections

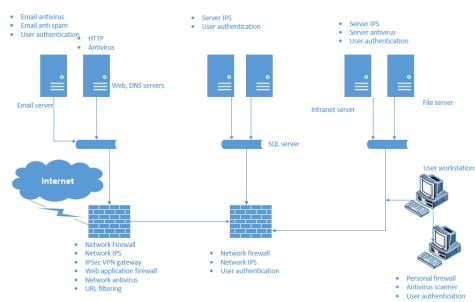


Fig. 2. Simulation results for the network.

and content inspection systems such as anti-virus, anti-spam, anti-malware and URL filtering. The security mechanisms in terms of operating systems database and applications are strengthened by hardware and software solutions. The strong security system built for large networks needs to be designed carefully after organization's risk analysis that fulfills security standards.

A. Security Rules

While designing network safety system, the main IT systems security principles that should be implemented are characterized below

1) Defence-in-depth: IT systems resources safety comprises of several security layers as illustrated in Fig 1. Defence-in-depth is followed by rules given below: Layered protection-number of security layers to amplify each other, what one misses the other catches.

Defence in multiple places network security system is placed at various locations of IT system. The security system of IT network resources depends on variety of safety layers that performs function of safeguard in different ways. Moreover, layers like two network firewalls having similarities must be originating from different vendors.

Meanwhile due to increased complexity of security system along with expensive management the implementation of this rule must be done carefully.

2) Compartmentalization of information: As shown in figure 2 given below for different sensitivity levels of IT network, there must be different security zones. In addition to this information hiding is also part of this rule that says IT system will provide only that data which is obligatory for carrying out IT system tasks (as only those servers will be registered in public DNS that are responsible for providing services to the internet).

3) Principle of least privilege: Another principle for IT system subjects is that they must be given with least benefits only required for proper functioning of organization. Same rule will be applied for external users using data and services. "Need-to-know" is extension of this rule which illustrates only that information will be provided to administrators of IT system that is necessary for them to perform their functions and duties.

4) Security zones: Firewall provides control on network traffic flow and it will flow according to command. Firewall

includes firewall devices, firewall functions in IPS devices and list of access control for system routers and switches. Firewalls play an important role in securing architecture by categorizing the IT system infrastructure into security zones having effective communication between them. This division rule is comprised of following points:

1) IT system resources of different sensitivity levels should be located in different security zones:

- Devices and computer systems providing services for external networks (e.g., the Internet) should be located in different zones (De-Militarized Zone - DMZ) than internal network devices and computer systems
- Strategic IT system resources should be located in dedicated security zones.
- Devices and computer systems of low trust level such as remote access servers and wireless network access points should be located in dedicated security zones.

2) IT system resources of different types should be located in separate security zones

- User workstations should be located in different security zones than servers
- Network and security management systems should be located in dedicated security zones
- Systems in development stage should be located in different zones than production systems

5) Intrusion prevention: IPS devices are responsible for detecting and blocking penetrations and attacks conducted by intruders and malicious malware applications. They should be installed in the network path between potential threat sources and sensitive IT system resources. When designing IPS systems, attacks through encrypted sessions (e.g., SSL) should also be taken into account. Since the IPS would not be able to inspect these encrypted sessions, an effective method would be to decrypt the sessions prior to IPS devices in order to inspect unencrypted packets. An important requirement for intrusion prevention tightness is the proper design of network protections and control rules. For one, internal networks should not have direct access to the Internet so a Trojan sent to a user's workstation through a phishing attack would not allow the intruder to connect to the external network.

IV. SECURITY IMPLEMENTATION

It is compulsory to develop a plan based to determine security risk for campus network. For this purpose, number of technologies like firewall technology, PKI technology, virtual private network (VPN) technology, virtual exchange network can be utilized to develop centralized configuration, and to strengthen monitoring management system. In short, formulating system must be made strong and specifications must be introduced to keep security confidential and implement it. The whole network system must be made secure in order to increase physical security of several equipment's of computer information system. It usually covers three features

that are lines security, environmental safety and equipment safety. Moreover, equipment safety covers several domains like anti-theft of devices, anti-electromagnetic information radiation leaks, anti-crash to secure power supply and anti-electromagnetic interference. Whereas environmental safety facilitates the security of system environment like disaster protection along with regional safety. Campus can be categorized into two parts that is internal network and external network. Internal network covers LAN, LAN office automation, library LAN and teaching LAN, whereas external network covers various public servers. According to network security viewpoint it is important to install internet firewall in the border along with implementation of security policy control. The internal network must be made secure by installation of probe for switcher on the main segment. This will work as intrusion detection system which keeps network access secure. As all computer terminals, servers and workstations are based on operating system to continue optimal function, so it is important to make secure operating system. Hence, genuine operating system with server version is appropriate for use in the critical servers and workstations such as database server, Email server, www server, backup server and network management stations, etc.

In the meantime, operating systems are at greater risk of insecurities because of inappropriate design and version, another reason behind security vulnerabilities is lack of knowledge about use of security settings. Without other higher security level commercial operating system for decision, working framework security administration and improvement of upgraded safety efforts is the key variables, furthermore, issues of operating system security may rise from virus threats and hackers penetrated the network to destroy the information. Therefore, anti-virus tools must have ability to secure network from every kind of virus attack through any internet source. In addition to, anti-virus program installation must be capable of removing virus program detected.

Hackers can hack the data for misuse and can make changes in system illegally. It is beneficial for system to install risk assessment programs so that administrators can easily recognize the threats and accept or reject user privileges command accordingly. Moreover, its fruitful to install real-time intrusion detection system IDS because it can trace out user's activity, and continuous check can secure internal staff from intranet destruction. Soon after detection of any harm, the system will immediately inform the administrator, then administrator is authorized to record the relevant test results and can use information to track hacker or intruder, and further can identify security risks to system and develop strategies to secure system's useful information and data from misuse or damage. Mostly student's computer room and campus network is connected to each other, and there is chance that virus can spread as students will definitely use the usb storage device on PC and students may frequently download some product from the Internet on the machine. Hence, to keep campus network safe from virus its important resolve the problem by using server antivirus, stand-alone anti-virus and

gateway anti-virus. Administrator can make network more secure by setting password and permissions in the network for every user, it not only saves user password and name but can also facilitate to keep user complete record with confidentiality. Network administrator must keep database of network users and system log in organized manner. To evaluate and audit the security circumstance frequently on the campus network, give careful consideration to the dynamic system security concerns and alter the significant security settings for interruption aversion, and recovery systems after emergency. Various security plans like network access control, the directory security control plan, access control policy, network monitoring, lock control and operation access control can be developed to make campus network security more strong. Another strategy i.e. information encryption strategy is beneficial to prevent unauthorized person from misuse or stealing of valuable data by applying encryption algorithm. There is various encryption programming which can encode messages; files et al. Utilize encryption programming can viably ensure information, documents, secret key, and control the data to transmit securely through the system. Back up and imaging technology can also be used to strengthen campus network. Data integrity can be increased by using backup technology, as data is stored in another place to make a backup, so data can be recovered easily once deleted from anywhere. When the development of campus network is being done, network security needs of a unified security system particular that ought to be combined with actual situation, and afterward the unified security system details ought to be actualized in the entire network. The administration ought to be executed first before the Implementation of network methods wellbeing, the development of the security association system is basic. Experts must be involved in designing and implementation of security plan and experts must coordinate with other departments for the implementation of security process and they should facilitate other departments with guidance to enhance the level of system's security.

The campus network data system can be made secure by taking technical measures, but besides this, management system should be improved and constructed in effective way. The campus network security is always at greater risk, so to expand usefulness for the security administrations, enhance network safety efforts, we should attempt awesome endeavours to fortify the system security administration, build up a secure administration framework and execute well secured administration standards, data arrange security is an exceptionally orderly work, just when the system security administration and the system security innovation are utilized at the same time, at that point they can develop a complete data organize security framework.

V. CONCLUSION

Recently number of colleges and schools develop their own particular IT systems, and the use of campus network turns out to be increasing generally, for this reason data security is an inescapable factor to guarantee the systems running

easily with maximum efficacy. Facilitating campus network with educational environment not only provides strong back up model but also promotes the information technique education. Maintenance of high efficiency of campus network is main problem that need to be resolved. This paper has highlighted various factors which can compromise campus security and give suggestions to sort out security related issues.

REFERENCES

- [Ali et al.(2015)] Ali, Hossain, and Parvez] Mohammed Nadir Bin Ali, Mohamed Emran Hossain, and Md Masud Parvez. Design and implementation of a secure campus network. *International Journal of Emerging Technology and Advanced Engineering*, 5:370–374, 2015.
- [Das and Behera(2017)] Kajaree Das and Rabi Narayan Behera. A survey on machine learning: Concept, algorithms and applications. 2017.
- [K C Ramakrishnegowda(2009)] B Shankarappa K C Ramakrishnegowda, U Kannappanavar. Campus network management: Best practice by kuvempu university. 2009.
- [MADJE(2017)] UMA MADJE. Design and analysis of network security model. pages 48–50, 2017.
- [Gursimrat Singh(2013)] Dr. Amardeep Singh Gursimrat Singh. Campus network security policies: Problems and its solutions. 2013.
- [Begum and Nandyal()] Faqarunnisa Begum and Suvarna Nandyal. Implementing different set of network security policies for a well infrastructure campus network.
- [Alabady(2009)] Salah Alabady. Design and implementation of a network security model for cooperative network. *Int. Arab J. e-Technol.*, 1(2): 26–36, 2009.

Improving the Data Quality in the Research Information Systems

Otmane Azeroual^{abc}

^a German Center for Higher Education Research

and Science Studies (DZHW), Schützenstraße 6a, Berlin, 10117, Germany

^b Otto-von-Guericke-University Magdeburg, Department of Computer Science, Institute for Technical and Business Information Systems Database Research Group, P.O. Box 4120; 39106 Magdeburg, Germany

^c University of Applied Sciences HTW Berlin, Study Program Computational Science and Engineering, Wilhelminenhofstraße 75A, 12459 Berlin, Germany
Azeroual@dzhw.eu

Mohammad Abuosba^c

^c University of Applied Sciences HTW Berlin, Study Program Computational Science and Engineering, Wilhelminenhofstraße 75A, 12459 Berlin, Germany
Mohammad.Abuosba@HTW-Berlin.de

Abstract — In order to introduce an integrated research information system, this will provide scientific institutions with the necessary information on research activities and research results in assured quality. Since data collection, duplication, missing values, incorrect formatting, inconsistencies, etc. can arise in the collection of research data in different research information systems, which can have a wide range of negative effects on data quality, the subject of data quality should be treated with better results. This paper examines the data quality problems in research information systems and presents the new techniques that enable organizations to improve their quality of research data.

Keywords — Research Information Systems, CRIS, RIS, Data Quality, Research Data, Data Cleaning, Data Profiling, Science System, Standardization

I. INTRODUCTION

With the introduction of a research information system, universities and non-university research institutions can provide an up-to-date overview of their research activities, record, process and manage information on their scientific activities, projects and publications, as well as integrate them into their web presence. For scientists, they offer opportunities to collect, categorize and use research information, be it for publication lists, for the preparation of projects, for the reduction of the effort required to produce reports, or for the external presentation of their research and scientific expertise. The data quality is of great importance. Only correct data can provide resilient, useful results and allow for a profound understanding of the research data of research establishment that are always up-to-date. The completeness, correctness and timeliness of the data are thus essential for successful operational processes. The data errors extend across different areas and weaken the entire research activities of an establishment. Therefore, the aim of this paper is to define and classify the problems of the quality of data that can occur in the research information systems, and then present new techniques that are used to recognize, quantify, resolve data

quality problems in research information systems, to improve their Data quality.

II. Research Information System (RIS) - Architecture

A RIS is a central database that can be used to collect, manage and provide information on research activities and research results. The information considered here provides metadata about research activities such as projects, third-party funds, patents, cooperation partners, prices and publications. The RIS architecture usually consists of (see figure 1):

- Data Access Layer
- Application Layer
- Presentation layer

The following figure provides an overview of the individual processes and shows which components belong to which process step:

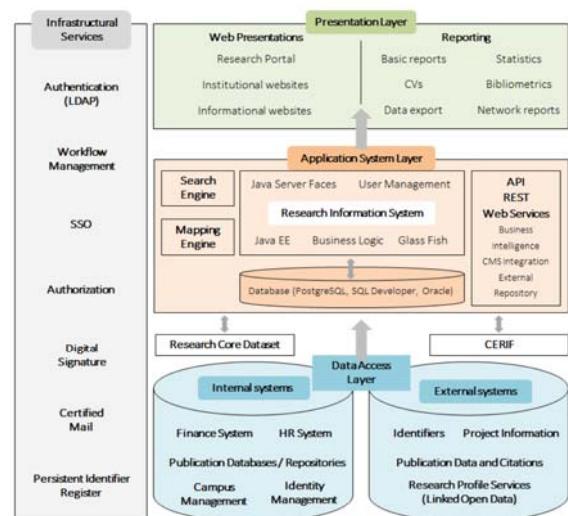


Figure 1: Architecture of RIS

The **data access layer** contains the internal and external data sources. This level contains, for example, databases from the administration or publication repositories of libraries, identifiers such as ORCID or bibliographic data from the Web of Science or Scopus etc. The **application layer** contains the research information system and its applications, which merge, manage and analyze the data held at the underlying level.

In the **presentation layer**, the target group-specific preparations and representations of the analysis results are depicted for the user. In addition to various possibilities of reporting, you can also fill portals and websites of the establishment here.

Orthogonal to the described layers, there are the Infrastructural Services, the overlapping services for the entire information systems, such as authentication, authorization, single sign on, etc.

Offers for the standardized collection, provision and exchange of research information in RIS are Research Core Dataset (RCD) and the Common European Research Information Format (CERIF) data model maintained by the non-profit organization euroCRIS Version CERIF 2008 1.0 is available. This data model describes the entities as well as their relationship to each other.

III. Problems of data quality

Collecting data in a database system is a standard process. At each facility, personal data, information about their scientific activities, projects and publications are entered and recorded. The processing and management of this data usually needs to be in a good quality, so that users can get better results.

The quality of data is often defined as the suitability of the data for use in certain required usage objectives, which must be error free, complete, correct, up to date and consistent. Requirements can be set by different stakeholders, in the RIS context e.g. especially by users of a RIS, but also by the RIS administrator. Poor quality data includes data errors, typographical errors, missing values, incorrect formatting, contradictions and further more. Such quality issues in RIS need to be analyzed and then remedied by data transformation and data cleansing. The following are the typical quality issues of data in the context of a RIS (see figure 2):

- Missing values (feature completeness)
- Incorrect information caused by input, measurement or processing errors (characteristic correctness)
- Duplicates in the dataset (feature redundancy-free)
- Unevenly represented data (feature uniformity)
- Logically contradictory values (consistency characteristic)

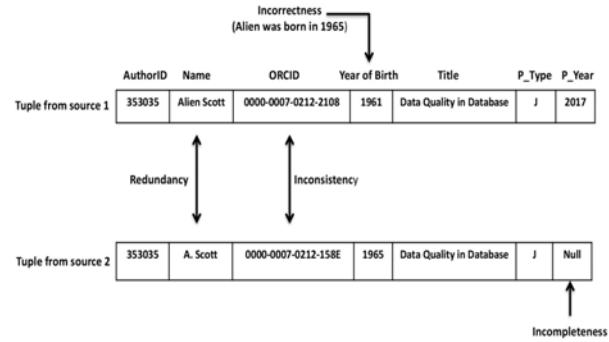


Figure 2: Examples of Data Quality Problems in RIS

IV. Improvement of data quality

Due to the integration of different internal data sources of the establishment and of external sources in research information systems, problems as stated in Chapter 3 have to be overcome. Now it is important to oppose the causes in this step and to improve the data quality.

The process of identifying and correcting errors and inconsistencies with the aim of increasing the quality of given data sources in RIS, is referred to as data cleansing (or "Data Cleaning") [8].

Data Cleansing includes all necessary activities to clean dirty data (incomplete, incorrect, not up to date, inconsistency, redundant, etc.). The data cleaning process can be roughly structured as follows [3]:

1. Defining and determining the actual problem
2. Find and identify faulty instances
3. Correction of the found errors

Data cleansing uses a variety of specialized methods and technologies within the data cleansing process. [9] subdivides them into the following phases (see figure 3):

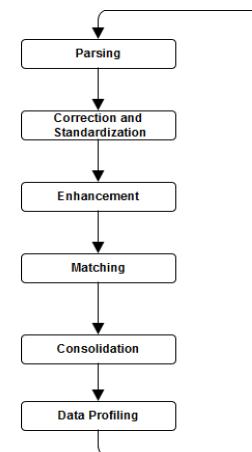


Figure 3: Data Cleaning Process

Parsing is the first critical component of data cleansing, helping the user to understand and transform the attributes more accurately. Individual data elements are referenced according to the metadata. This process locates, identifies and isolates individual data elements. For this process, e.g. for names, addresses, zip code and city. The parser Profiler analyzes the attributes and generates a list of tokens from them, and with these tokens the input data can be analyzed to create application-specific rules. The biggest problems here are different field formats that have to be detected.

Correction & Standardization is further necessary to check the parsed data for correctness, and then correct it afterwards. Standardization is the prerequisite for successful matching and there is no way around using a second reliable data source. For address data, a postal validation is recommended.

Enhancement or data enrichment is the process that expands existing data with data from other sources. Here additional data is added to close existing information gaps. Typical enrichment values are demographic, geographic or address information.

Matching: There are different types of matching: for reduplicating, matching to different datasets, consolidating or grouping. The adaptation enables the recognition of the same data. For example, redundancies can be detected and condensed for further information.

Consolidation (merge): Matching data items with contexts are recognized by bringing them together.

Data profiling is what the data is to be understood, i. E. capture technical structures, analyze them for the purpose of detecting data quality issues, and identify business inconsistencies. Profilers do not describe business rules and do not make any changes. They are only for analyzing the data. Profilers are often used at the beginning of a data analysis, but can also help to better illustrate the results of the analysis.

All of these steps are essential for achieving and maintaining maximum data quality in research information systems. Errors in the integration of multiple data sources in a RIS are eliminated by the clearing up.

The following Table 1 illustrates an example of identifying records with faulty names in a publication list to show how data cleansing processes (clearing up, standardization, enrichment, matching, and merging) can improve the quality of data sources.

The clearing up process adds missing entries, and completed fields are automatically adjusted to a specific format according to set rules.

Original Data before clearing up

Data Source					
Author ID	Name	ORCID	Birth Date	Address	
353035	Alien Scott	0000-0007-0212-2108	10/25/1965	145 F. Concord Street, Orlando, 32801	
353035	Dr. Alien Scott	0000-0000-0000-0000	25.10.1965	Concord Street, 32801 145F	
353035	Alien William Scott	0000-0007-0212-2108	652510	25 Concord 32801 Street	
353036	A. Scott		11/25/56	12 Ford Ave 32801	
353036	Scott Allen	0000-0007-0212-2108	25.11.1965		
	Alien Scott	0000000702122108	25.10.1956	Street C, 32801 145F. US	
410003	Olivia Svenson	0450-1254-3598-F156	6-2-1983	745-7801 P.B. Las Vegas 29502	
410003	Svenson Olivia	045012543598F156	1983	7801 P.B. Las Vegas 29502	

Data after clearing up

In this example, the missing zip code is determined based on the addresses and added as a separate field. Enrichment rounds off the content by comparing the information against external content, such as demographic and geographic factors, and dynamically expanding and optimizing it with attributes.

Cleansed Data					
Author ID	First	Last	ORCID	Birth Date	Address
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	32801; FL; Orlando; 145 F. Concord Street
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	32801; FL; Orlando; 145 F. Concord Street
353035	Alien	Scott	0000-0007-0212-2108	1965-11-25	32801; FL; Orlando; 145 F. Concord Street
353036	Alien	Scott	0000-0007-0212-2108	1956-11-25	32801; FL; Orlando; 12 Ford Ave
353036	Alien	Scott	0000-0007-0212-2108	1965-10-25	FL; Orlando; 145 F. Concord Street
410003	Olivia	Svenson	0450-1254-3598-F156	1971-02-06	29502; NV; Las Vegas; 745-7801 PO Box
410003	Olivia	Svenson	0450-1254-3598-F156		29502; NV; Las Vegas; 745-7801 PO Box

Data before enrichment

Cleansed Data					
Author ID	First	Last	ORCID	Birth Date	Address
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	FL; Orlando; 145 F. Concord Street
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	FL; Orlando; 145 F. Concord Street
353035	Alien	Scott	0000-0007-0212-2108	1965-11-25	FL; Orlando; 145 F. Concord Street
353036	Alien	Scott	0000-0007-0212-2108	1956-11-25	FL; Orlando; 12 Ford Ave
353036	Alien	Scott	0000-0007-0212-2108	1965-10-25	FL; Orlando; 145 F. Concord Street
410003	Olivia	Svenson	0450-1254-3598-F156	1971-02-06	NV; Las Vegas; 745-7801 PO Box
410003	Olivia	Svenson	0450-1254-3598-F156		NV; Las Vegas; 745-7801 PO Box

Data after enrichment

The example shows how the reconciliation and merge process runs. Merging and matching promote consistency because related entries within a system or across systems can be automatically recognized and then linked, tuned, or merged.

Enriched Data					
Author ID	First	Last	ORCID	Birth Date	Address
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	FL; Orlando; 145 F. Concord Street 32801
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	FL; Orlando; 145 F. Concord Street 32801
353035	Alien	Scott	0000-0007-0212-2108	1965-11-25	FL; Orlando; 145 F. Concord Street 32801
353036	Alien	Scott	0000-0007-0212-2108	1956-11-25	FL; Orlando; 12 Ford Ave 32801
353036	Alien	Scott	0000-0007-0212-2108	1965-10-25	FL; Orlando; 145 F. Concord Street 32801
410003	Olivia	Svenson	0450-1254-3598-F156	1971-02-06	NV; Las Vegas; 745-7801 PO Box 29502
410003	Olivia	Svenson	0450-1254-3598-F156		NV; Las Vegas; 745-7801 PO Box 29502

Matching

This example finds related entries for Alien Scott and Olivia Svenson. Despite the similarities between the datasets, not all information is redundant. The adjustment functions evaluate the data in the individual records in detail and determine which ones are redundant and which ones are independent.

Cleansed Data					
Author ID	First	Last	ORCID	Birth Date	Address
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	32801; FL; Orlando; 145 F. Concord Street
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	32801; FL; Orlando; 145 F. Concord Street
353035	Alien	Scott	0000-0007-0212-2108	1965-11-25	32801; FL; Orlando; 145 F. Concord Street
353036	Alien	Scott	0000-0007-0212-2108	1956-11-25	32801; FL; Orlando; 12 Ford Ave
353036	Alien	Scott	0000-0007-0212-2108	1965-10-25	32801; FL; Orlando; 145 F. Concord Street
410003	Olivia	Svenson	0450-1254-3598-F156	1971-02-06	29502; NV; Las Vegas; 745-7801 PO Box
410003	Olivia	Svenson	0450-1254-3598-F156		29502; NV; Las Vegas; 745-7801 PO Box

Consolidation

The merge makes the reconciled data a comprehensive data set. In this example, the duplicate entries for Alien Scott are merged into a complete record containing all the information.

Cleansed Data					
Author ID	First	Last	ORCID	Birth Date	Address
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	32801; FL; Orlando; 145 F. Concord Street
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	32801; FL; Orlando; 145 F. Concord Street
353035	Alien	Scott	0000-0007-0212-2108	1965-11-25	32801; FL; Orlando; 145 F. Concord Street
353036	Alien	Scott	0000-0007-0212-2108	1956-11-25	32801; FL; Orlando; 12 Ford Ave
353036	Alien	Scott	0000-0007-0212-2108	1965-10-25	32801; FL; Orlando; 145 F. Concord Street
410003	Olivia	Svenson	0450-1254-3598-F156	1971-02-06	29502; NV; Las Vegas; 745-7801 PO Box
410003	Olivia	Svenson	0450-1254-3598-F156		29502; NV; Las Vegas; 745-7801 PO Box

Golden Record					
Author ID	First	Last	ORCID	Birth Date	Address
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	32801; FL; Orlando; 145 F. Concord Street
410003	Olivia	Svenson	0450-1254-3598-F156	1971-02-06	29502; NV; Las Vegas; 745-7801 PO Box

For the frontend of the RIS could be checked with the profiling process, when it comes to the production of statistics,

reports on research data. Profiling makes it easy to assess the overall condition of the data, to identify, prioritize and correct errors, and to remedy the cause of quality issues. Once a profile is created, a facility can respond to quality issues by continually monitoring profile-related parameters.

V. Discussion

The clearing up of data errors in RIS is one of the possible ways to improve existing data quality. Following the defined data cleansing processes, the following developed use cases could be identified in the target system and should serve as a model to show how to detect, quantify, correct and improve them in the case of data errors in research information systems in the establishment.

The following figure introduces the just-mentioned use cases for improving data quality in RIS.

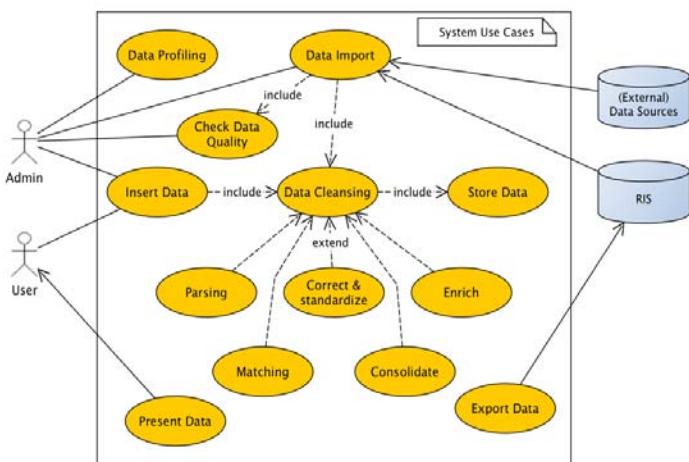


Figure 4: Use Case for Improving Data Quality in the RIS

The meta process flow can be viewed as shown in the following figure:

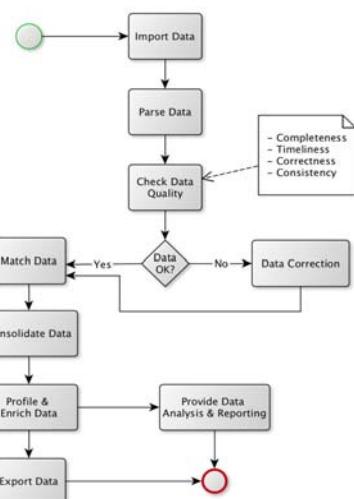


Figure 5: Main Workflow of the Process

The techniques presented by Data Cleaning help establishments that overcome problems. With these steps every establishment can successfully enforce their data quality.

VI. Conclusion and Outlook

In this paper, the question was addressed, which quality problems can occur in research information systems and how to fix and improve them with new techniques or methods, such as Data Cleansing.

As a result, it were been shown that the improvement of the data quality can be performed at different stages of the data cleansing process in any research information system and that high data quality can be obtained from universities and research institutions to operate e.g. research information systems successfully. In this respect, the review and improvement of data quality are always targeted. The illustrated concept can be used as a basis for the using facilities. It offers a suitable procedure and a use case, on the one hand to be able to better evaluate the data quality in RIS, to be able to prioritize problems better and to prevent them in the future as soon as they occur. On the other hand, these data errors must be corrected and improved with Data Cleansing. It says: "The sooner quality defects are detected and remedied, the better." Already in the acquisition phase, the author himself or a downstream control authority can correct software errors, such as typing errors, missing values, incorrect formatting, contradictions, etc. To support the universities and all research institutions in the implementation of the data cleansing process, there are numerous tools. With these tools, all facilities can significantly increase the completeness, correctness, timeliness, and consistency of their key data, and they can successfully implement and enforce formal data quality policies.

Data-Cleansing tools are primarily commercial and available for both small application contexts and large data integration application suites. In recent years, a market for data cleansing is developing as a service.

In future work, possible expert interviews and / or quantitative surveys are planned with the research facilities and universities in order to find out how high the data quality is in their research information system and what measures to improve the quality of data are applied to research information.

REFERENCES

- [1] Batini, C. and Scannapieco, M. Data Quality - Concepts, Methodologies and Techniques; Springer-Verlag, Heidelberg, 2006.
- [2] Gebauer, M. and Windheuser, U. Structured Data Analysis, Profiling and business rules, Springer Fachmedien Wiesbaden 2015.
- [3] Helmis, S. and Hollmann, R. Web-based data integration, approaches to measuring and securing the quality of information in heterogeneous data sets using a fully web-based tool, 1. edition © Vieweg+Teubner | GWV Fachverlage GmbH, Wiesbaden 2009.

- [4] Hildebrand, K., Gebauer, M., Hinrichs, H. and Mielke, M. Data and information quality. On the way to the information excellence, 3rd, extended edition, Springer Fachmedien Wiesbaden 2015.
- [5] Martin, M. Measuring and Improving Data Quality. Part II: Measuring Data Quality. In: NAHSS Outlook. Ausgabe 5, April 2005.
- [6] Müller, H. and Freytag, J.-C. Problems, Methods and Challenges in Comprehensive Data Cleansing, Technical Report HUB-IB- 164; Humboldt Universität, Berlin, 2003.
- [7] Naumann, F. and Rolker, C. Assessment Methods for information Quality Criteria. In: Proceedings of the MIT Conference on Information Quality, pp. 148-162.
- [8] Naumann, F. and Leser, U. Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen. Dpunkt Verlag, 1. Edition, Oktober 2007.
- [9] Rahm, E. and Do, H.H. Data cleaning: Problems and current approaches. IEEE Bulletin of the Technical Committee on Data Engineering, 23(4), 2000.

AUTHORS PROFILE



Otmane Azeroual, is working as a researcher in the project "Helpdesk for the Introduction of the Research Core Dataset" at the Department 2 Research System and Science Dynamics, German Center for Higher Education Research and Science Studies (DZHW). After studying Business Information Systems at the University of Applied Sciences Berlin (HTW), he began his PhD in Computer Science at the Institute for Technical and Business Information



Prof Dr.-Ing. Mohammad Abuosba, is a Professor at the Department of Study Program Computational Science and Engineering, University of Applied Sciences (HTW) Berlin. His Research areas are Engineering, IT Systems Engineering (focus on Database Systems, Product Data Management), Modeling, Quality Management and Project Management.

One hop forwarding technique for QOS routing with an Improved Fault tolerant Clustering Mechanism for Multimedia Sensor Networks

¹ R. Guru, ² Dr. Siddaraju, ³ Ananda babu

¹Research scholar, ²Professor, ³Assistant Professor

¹Sri Jayachamarajendra College of Engineering, Mysuru, India.

²Dr. Ambedkar Institute of Technology, Bangalore, India

³Kalpataru Institute of Engineering, Tiptur, India.

¹guruirg@sjce.ac.in, ² siddaraju_b@yahoo.co.in, ³babu.tiptur@gmail.com

Abstract: Wireless Sensor Network (WSN) is a wireless network consists of large number of sensor nodes. Routing protocols are in charge of discovering and maintaining the routes in the network. Finding the shortest path is one of challenging factor which greatly enhances the performance of the network. The first section of this paper focuses on finding best possible next hop to route the packets to reach the sink node using TPGF protocol. Next section of this paper concentrate on multimedia applications where they have taken an essential role in our daily lives, and their usage is growing day by day. The multimedia sensors have the ability to capture video, image, audio, and scalar sensor data and deliver the multimedia content through sensors network. However, due to the reason of their location and size of data they transmit for multimedia may cause the system failure in many cases. To overcome this paper presents an Improved Fault tolerant distributed clustering mechanism (IFCMM). This mechanism use run time for the recovery of sensor nodes due to the failure of head of cluster nodes. Experimental results show an improvement in the performance of the algorithm with respect to various metrics of QOS.

Keywords: Geographic routing, Transmission of multipath, Realistic conditions of sensor.

I. INTRODUCTION

Effectively transmitting sight and sound streams in remote interactive media sensor systems (WMSNs) is a noteworthy testing issue, because of the constrained transmission transfer speed and control asset of sensor hubs. Three late reviews [1–3] on interactive media correspondence in wireless multimedia sensor networks which demonstrates that present conventions in both interactive media and sensor systems fields that are not appropriate for both sight and sound correspondence in WMSNs, in light of the fact that they don't have enough thought on the attributes of sight and sound gushing information also, characteristic obliges of sensor systems in the meantime. These three overviews additionally expounded that there is no arrangement concentrating on tending to the steering issue of sight and sound spilling in geographic WMSNs.

1.1 Transmission of Multipath: Packets of sight and sound spilling information for the most part are vast in size and the transmission prerequisites can be a few times higher than the most extreme transmission limit (data transfer capacity) of sensor hubs. This requires that multipath transmission ought to be utilized to increment transmission execution in WSNs [4].

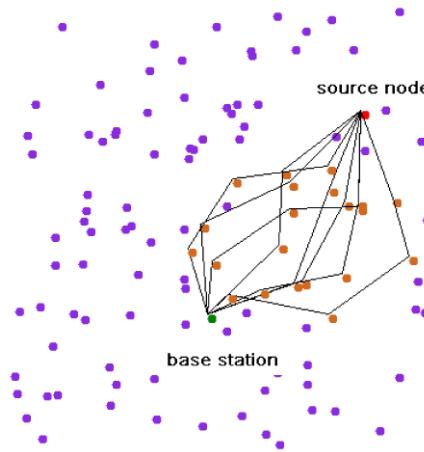


Figure 1: Dynamic Hole: Constructed by a group of sensor nodes present in the present routing paths

1.2 Bypassing the Hole: This includes how the openings of dynamic may happen due to sensor hubs which are little in range over-burden because of the sight and sound transmission, Figure 1 describes how proficiently bypassing these element openings is an important part of transmission in WSNs.

1.3 Shortest way transmission: Multimedia applications by and large have a postpone imperative, which requires that the sight and sound spilling in WSNs ought to dependably utilize the most limited steering way, which has the base end-to-end transmission delay.

Media transmission in WMSNs requires another directing calculation that can bolster all these three necessities in the meantime. This paper proposes another Two-Phase geographic Greedy Forwarding (TPGF) directing calculation for investigating one or numerous most limited (close briefest) hole bypassing transmission ways in WMSNs. The principal period of TPGF is in charge of investigating the conceivable directing way. The second period of TPGF is in charge of improving the discovered steering way with minimal number of bounces. TPGF can be executed over and again to discover various on-request node disjoint steering ways. TPGF has the accompanying essential elements that make it be not quite the same as existing geographic steering calculations [5–8].

TPGF is an immaculate geographic voracious sending directing calculation. It does exclude the face steering idea, e.g., right/left hand guidelines and tally/clockwise edges, which is not the same as many existing geographic sending directing calculations, e.g., GSPR [5].

TPGF does not require the calculation and protection of the planar diagram in WMSNs. This point permits more connections to be accessible for TPGF to investigate more node disjoint steering ways, since utilizing the planarization calculations really restrains the useable connections for investigating conceivable steering ways.

TPGF does not have the outstanding Local Minimum Issue [5], which is characterized as "a sensor hub finds no next-bounce hub that is nearer to the base station than itself". Explore work in this paper has made both hypothetical also, viable commitments to comprehend the geographic directing in WMSNs. The hypothetical commitments are: It is demonstrated that: there exists a geographic covetous sending directing calculation (TPGF) that can ensure bundle conveyance (bypassing openings) in any 2D/3D sensor systems without utilizing the face directing strategy, when sensor hubs just think about their 1-bounce neighbor hubs.

It is demonstrated that: there exists a geographic avaricious sending directing calculation (TPGF) that can locate the most limited directing way (or close most limited steering way when openings exist) for limiting the end-to-end transmission delay, at the point when the gaps data is not recognized ahead of time. The handy commitments in this paper are as taking after four viewpoints:

- **Key curiosity:** To the best of our insight, TPGF is the to start with unadulterated geographic avaricious sending steering calculation that spotlights on supporting interactive media spilling in WMSNs, which bolsters the accompanying three elements at the same time.
- **Supporting multipath transmission:** TPGF can discover one directing way per execution and can be executed over and again to discover more on-request hub disjoint directing ways.
- **Supporting gap bypassing:** TPGF gives a superior arrangement for gap bypassing in both 2D and 3D sensor systems than other related research work.
- **Supporting most brief way transmission:** TPGF can discover the most brief steering way (or close most limited directing way when openings exist) for limiting the end-to-end transmission delay.

TPGF controlling computation can make a tremendous effect on both adaptable sight and sound and remote sensor frameworks (WSNs) research gatherings.

Remote sensor systems have increased tremendous consideration for their extensive variety of utilizations, for example, natural checking, military reconnaissance, human services, and debacle administration [1]. One of the real imperatives of WSNs is the constrained and by and large indispensable power wellsprings of the sensor hubs. Undoubtedly, in various applications, it is impossible to supplant the sensor center points as they work under savage condition. In this way, diminishing imperativeness usage of the sensor centers is considered as the most essential test for long run operation of WSNs. Expansive looks at have been finished in arranging essentialness saving traditions which consolidate low-control radio correspondence hardware, imperativeness careful MAC traditions, et cetera.

In any case, vitality productive grouping and steering calculations [2, 3] are the most two promising territories that have been contemplated widely for WSNs. In a bunch based WSN (allude Figure . 1), the sensor hubs are composed into unmistakable gatherings, called groups. Each gathering has a pioneer, called group head (CH) and every sensor hub has a place with one and just a single bunch. Bunching WSN has following points of interest. (1) It empowers information conglomeration at group make a beeline for dispose of the repetitive and uncorrelated information, subsequently diminishing vitality utilization of the sensor hubs. (2) Routing can be all the more effectively oversaw on the grounds that lone CHs need to keep up the neighborhood course setup of different CHs and hence requiring little directing data. This thusly enhances the versatility of the system essentially. (3) It likewise moderates correspondence transmission capacity as the sensor hubs speak with their CHs just and in this way stay away from the trading of repetitive messages among themselves. Be that as it may, in grouping approach, a CH bears some additional work stack, i.e., getting detected information sent by part sensor hubs, information total and information scattering to the BS.

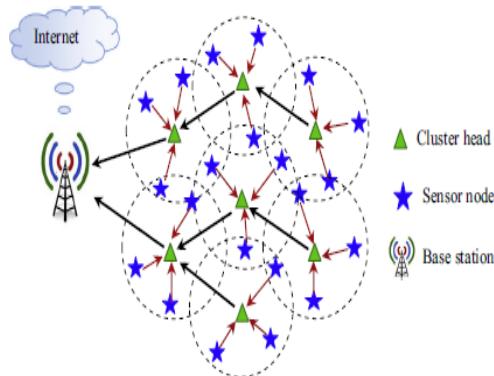


Figure 2: Model of wireless sensor network

Additionally, in numerous WSNs, the CHs are typically chosen among the ordinary sensor hubs, which can bite the dust rapidly as they devour more vitality because of such additional work stack. In this specific situation, numerous analysts [4– 10] have proposed the utilization of some exceptional hub called entryways or transfer hubs that are provisioned with additional vitality. These passages are dealt with as the bunch heads (CHs) which are in charge of a similar usefulness of the CHs. Sadly the entryways are likewise battery worked and thus control compelled.

In other grouping calculations, the bunching is framed by picking chose number of group heads and changing the CHs on each cycle. Be that as it may, the calculation does not consider if the chose CH comes up short amid its cycle. This may prompt disappointment of parcels because of system soften up the framework. Subsequently, the proposed framework thinks about disappointment case.

Intellectual nodes are most promising because of innovation in the WSNs [1]. It bolsters extremely introverted correspondence worldview suits remote sensor organizes that convey well proportioned movement because of their sudden behavior. These intellectual sensor nodes can help defeat the issues of failure of sensor nodes and these nodes will exorbitant dispute in WSNs that emerge because of the sending of a few sensors associated through radio connections. In correlation, ordinary WSNs utilize settled range allotment approach and their execution is restricted because of low preparing and correspondence energy of sensor hubs which are ordinarily asset compelled. In any case, WSNs work over unlicensed groups which are getting to be plainly immersed because of a huge development in the quantity of remote applications utilizing similar groups. Subsequently, the test is to proficiently use the range and this test can be tended to by utilizing psychological radio innovation for WSNs.

This paper also considers video perception based on different condition for blended sensor nodes in the network. Scholarly the framework this sensor nodes include the circumstance, which intuitive media sensor that centers outfitted with cameras is related with remote associations in exceptionally designated way. No less than one of these center points is dynamic meanwhile. They endlessly screen the earth and transmit the got accounts to the sink center point. These center points take a shot at batteries and accordingly imperativeness capability is a fundamental issue for this structure. Likewise, for remote transmission of video the range should be utilized successfully as nowadays run in urban regions is an uncommon resource. The approved range gatherings (3G, 4G, et cetera.) are exceedingly utilized and the unlicensed range gatherings, for instance, ISM bunches are winding up perceptibly exceptionally drenched. That is the reason we propose an imperativeness profitable approach in light of subjective radio for a successful utilization of the range.

The key issues and challenges of MWSN are high transmission limit ask for, high essentialness use, nature of organization (QoS) provisioning, data taking care of, and compacting frameworks, and cross-layer design:

- MWSN requires high transmission limit with a particular true objective to pass on media substance, for instance, a video stream, sound stream or pictures.

Multimedia transmission requires certain QoS guarantees. Regardless, QoS provisioning is to a great degree testing in WSNs as radio associations can have variable breaking point and deferral.

Energy capability is fundamental for MWSN as sensor centers frequently have low essentialness resources.

CR innovation can answer the above prerequisites of MWSNs. CR can give additional data transmission and enhance the nature of administration. Be that as it may, such intellectual radio methodologies, particular to MWSNs, should be outlined which concentrate on vitality productive correspondence to abuse transmission power and range qualities versus execution and unwavering quality exchange off. Moreover, minimal effort calculations should be intended for range detecting and dynamic range use. A large portion of the works in the writing concentrating on CRSNs are identified with just range detecting [20]-[22]. A couple of works concentrate on sight and sound transmission over intellectual radio systems [23]-[26], however they don't consider the WSN condition and the related limitations. This paper presents CR technology for MWSNs and the key contributions of this paper are:

- The spectrum aware based clustering mechanism for CRWSN: The clustering mechanism is aware of spectrum and takes it considers the accepted state of channels.
- This paper also includes head election mechanism for energy aware cluster: The proposed mechanism also considers the residual energy of the candidate cluster head

And energy consumed by the elected node.

II. RELATED WORKS

Some present works related to CR [1] [2], focus on go identifying, dynamic range get to, MAC traditions [3], controlling traditions and QoS [4]. Above is just a reviewing of the composition and various diverse works exists for CR for improvised frameworks and cell frameworks since it is particularly packed in composing. In any case, the examination on applying CR to WSNs is still in its underlying stage. The examination game plans proposed for all around helpful CR frameworks can't be particularly associated with CRSNs by virtue of the remarkable components of WSNs, for instance, the confined resources and imperativeness restrictions. Thusly, more a la mode CR plans are required that should be changed for WSNs to speak to resource and essentialness goals. If we especially focus on MWSNs by then there are various essentialness beneficial techniques, as assembled in a diagram on MWSN [5] [6], however mental radio isn't considered is those procedures. In particular, using CR with MWSN incorporates a tradeoff between go availability, QoS, essentialness usage and resource capability.

A couple of works related to blended media spilling in CRSNs are delineated in the going with content. The going with works conveys issues related to different layers, for instance, transport, framework, MAC and PHY layers. Bicen et al. [7] discuss the primary arrangement challenges for sight and sound and deferral delicate data transport in CRSN. They research particular transport layer traditions from the point of view of sight and sound transport in

CRSNs. In any case, unique issues like packing and guiding are not taken care of. Remembering the ultimate objective to address the issue of high information exchange limit solicitations and QoS necessities, Rehmani et al. [8] proposed channel holding for CRSNs. The thinking is to bond the flanking channels when appeared differently in relation to channel add up to which can add up to non-neighboring channels, yet is more personality boggling for WSNs. They proposed the channel holding thought and called consideration an excessive number of issues and troubles. In any case, the computations and traditions to perform perfect channel holding were not viewed as and were left for future work.

Liang et al. [9] thought about the QoS execution in CRSNs. They propose a need based framework in MAC layer. The steady development is arranged in guaranteed plan openings and the best effort action is served in strife get to period. They mull over the execution using investigative models and exhibit that appealing execution can be expert for consistent development. An expansion of this examination is presented in [10]. In any case, they don't consider the imperativeness use and simply the case related to one gathering of sensor centers are inspected. Multi-gatherings, batching instruments and bury cluster correspondences are not considered.

Gathering is considered by Fard et al. [11], who propose a QoS MAC tradition for remote sight and sound sensor frameworks using multi-channel approach. The remote development is requested into different classes and assorted channels are adaptively named to different action. This approach upgrades the throughput and delays 4 displays while being imperativeness capable. Regardless, PU development isn't contemplated and each cluster head is acknowledged to have various radio interfaces, which isn't useful. Wen et al. [12] propose a QoS coordinating tradition for extraordinarily named CRNs by considering parameters, for instance, open time, repeat gatherings, transmission run, botch rate, basic customer (PU) impedance rate and transmission run. Nevertheless, they simply focus on end-to-end delay.

Zhou et al. [13] propose a scattered booking figuring for video spouting over multi-channel multi-radio and multi-bounce remote frameworks. This approach makes appropriated booking designs with the objective of constraining the video mutilation, considering conventionality. The booking is characterized as a bended change issue, and an answer is proposed by together considering channel errand, rate circulation, and controlling. The work considers the tradeoff between constraining video twisting of each stream as opposed to upgrading overall execution by restricting framework blockage. Nevertheless, they don't consider subjective radio. They acknowledge all channels are open and dynamic range get to considering PU development isn't considered. What's more, observe that imperativeness organization isn't considered in the above works.

Low et al. [24] have proposed a figuring which uses a bipartite chart of the sensor center points and the entryways for finding a most outrageous planning of naming a sensor center to a CH. The figuring has space plan savvy multifaceted nature of $O(mn^2)$ for n sensor center points and m CHs. This is high for an immense scale WSN. It in like manner requires building a BFS tree for an individual sensor center point which takes a liberal measure of memory space. In [26] creators have proposed a figuring with $O(n\log n)$ time, which is a change over [6]. Regardless, both the figuring's having not considered the essentialness use issue. In [10], an Energy Efficient Load-Balanced Clustering Algorithm (EELBCA) has been proposed with $O(n\log n)$ time. EELBCA watches out for imperativeness adequacy and likewise stack changing. Low-vitality versatile bunching chain of importance

(LEACH) [29] is a well known procedure that structures groups by utilizing a conveyed calculation. It progressively pivots the work heap of the CH among the sensor hubs, which is valuable for stack adjusting. Be that as it may, the primary burden of this approach is that a hub with low vitality might be chosen as a CH which may pass on rapidly. In MHRM (least bounce directing model) [30], each hand-off hub finds a way to the BS such that the jump check is limited. Hence, much measure of vitality is spent to transmit information as it chooses farthest transfer hub in its correspondence go. In [26], creators have built up a disseminated blame tolerant system called CMATO (Cluster-Member-based blame Tolerant component) for WSNs. The creators use the catching methods to recognize the blame. At whatever point a CH falls flat, group individuals identify it and re-select another CH to shape another bunch.

III. PROPOSED ALGORITHM

This paper presents the Two-Phase Geographic Greedy Forwarding (TPGF) routing algorithm for which explores one or more multiple shortest hole by transmitting the paths in wireless multimedia sensor networks, the paper introduces clustering algorithm with energy efficient fault tolerant mechanism. Protocol uses Greedy forwarding technique. A forwarding node finds its next hop node which is closer to the base station using shortest distance vector path. This resolves finding shortest route. Shortest path always leads to depletion of energy of nodes falling in the path. This accounts for creation of dynamic hole. To overcome this, step back rule is introduced. Alternative path is searched through reverse path to ensure the packet will not be dropped due to dynamic hole formation.

- TPGF mechanism is used to convert the WSN to Distance based Search Tree (DST) using the routing algorithms presented in GPSR [5].
- While converting the nodes into TPGF_DST, all neighbor nodes will be added to the search tree.
- Path circle may be formed if two or more nodes within the path are neighbor nodes of the sensor node fall in the path.
- TPGF eliminates the path circle problem by implementing label based path optimization. Path optimization eliminates unnecessary nodes which are part of the path formed.
- Source node establishes a path to destination node at beginning. The response message is received all through from destination node with label number of each node that participated. If any node finds two of its neighbors involved in path formation, then the farthest node from base station will be eliminated from path. This improves the performance of the network.

This paper also presents an Improved Fault Tolerant Clustering Mechanism for Multimedia Transmission of Wireless Sensor Networks (**IFCMM** WSNs). This mechanism described in two different ways:

3.1 Clustering system

- Network is divided into three types of nodes: CH, non-CH and gateways
- Gateways are selectively chosen based on their location
- Gateways play major role while transmitting packets

- In case of CH failover, nonCH nodes within the network will apply fault tolerance mechanism. They initiate a broadcast request to know neighbor CH information. Nearby gateway upon receiving the request will initiate them as CH and respond back to the requested nonCH node.
- Gateways broadcasts its change of role to all nearby nodes and nodes will update in their memory

3.2 Multimedia

- All nodes are operated in multi channels. And nodes are divided into Primary User (PU) and Secondary User(SU)
- PUs will start their transmission by default
- SUs will look for channel availability by sensing the channel
- Intra-cluster communication will be achieved by TDMA algorithm
- Inter-cluster communication will be achieved by CDMA algorithm by SUs

3.3 Distributed clustering algorithm

There are many clustering protocols developed so far. This paper extends the clustering to a next level by introducing gateways. Gateways are special nodes deployed selectively within the network. These play a substitute role when CH node dies. When no response is received by nonCH node, issues broadcast message. Gateways nearby hearing the message will initiate the role of CH. Figure 3.1 explains distributed fault tolerant algorithm.

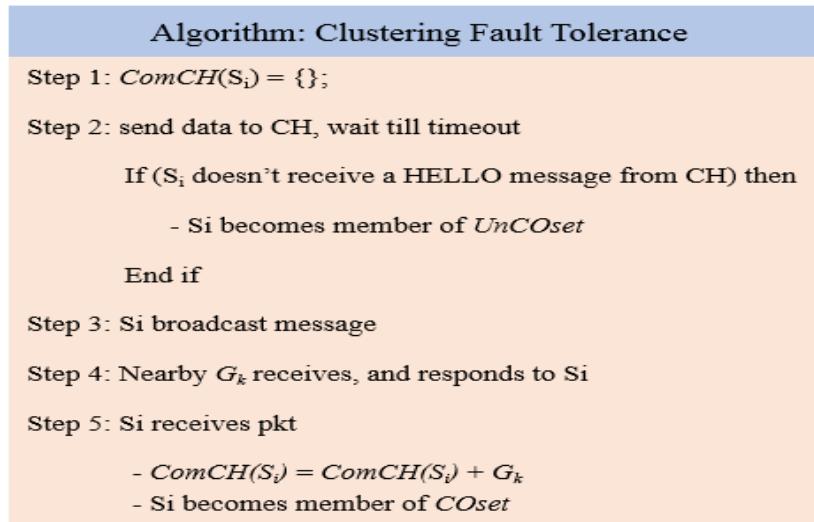


Figure 3.1: Distributed Fault Tolerant Clustering Algorithm

VI. Performance Analysis

In all the experiments, we have randomly deployed the nodes within the network. In the first phase, the protocol is implemented without optimizing the transmission path. Fault tolerance modal is implemented to evaluate the performance of the protocol. Delivery of the packets is calculated against distinct number of nodes taken into consideration. Results are compared with GPSR protocol. Proposed protocol shows a good result with respect to GPSR in terms of packet delivery ratio as shown in the figure 4.1 and 4.2 with fault tolerance value of 10% and 20% respectively.

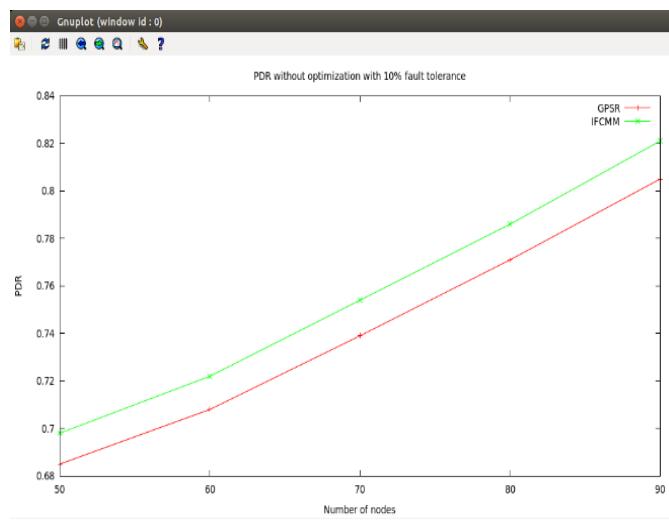


Figure 4.1: PDR without path optimization F=10%

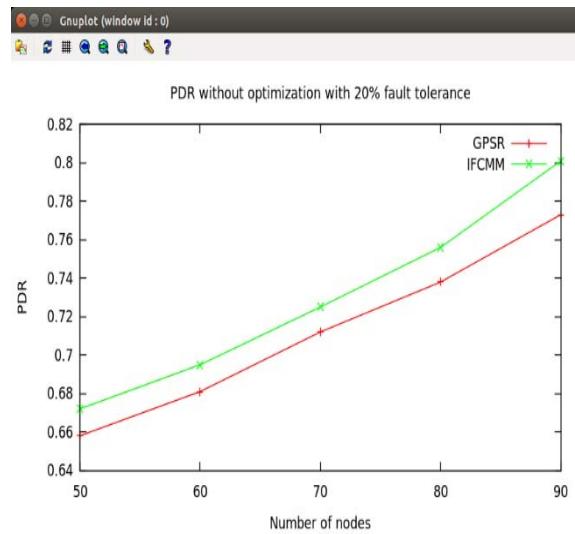


Figure 4.2: PDR without path optimization F=20%

Also, the results are calculated with path optimization. Figure 4.3 and 4.4 shows packet delivery ratio results with fault tolerance of 10% and 20% respectively. In GPSR, the routing is just to the next hop of the node from which the packet is being sent. If the receiving node doesn't have next forwarding node, then the packet will be lost. And this causes the degraded performance. Whereas IFCMM uses greedy forwarding with step-back technique to find a different route and ensure a good result in delivering the packets.

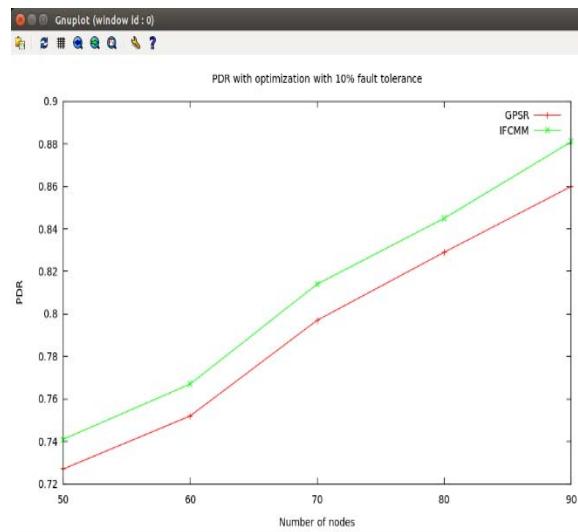


Figure 4.3: PDR with path optimization F=10%

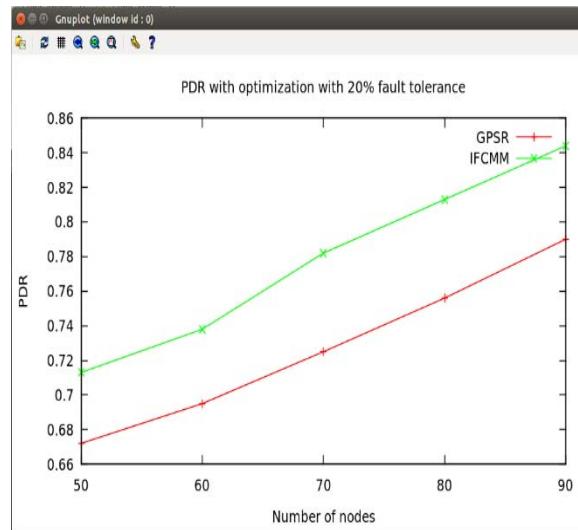


Figure 4.4: PDR with path optimization F=20%

The results are calculated for the energy consumption against the number of data packets delivered. Figure 4.5 shows the energy consumption results plotted against Direct Diffusion protocol. The results show the energy consumed by the system is better than DD, since IFCMM uses the path optimization with fault detection. In DD, the location identification of nodes is through receiving signal strength and this may cause inaccurate identification and may result in disjoint path.

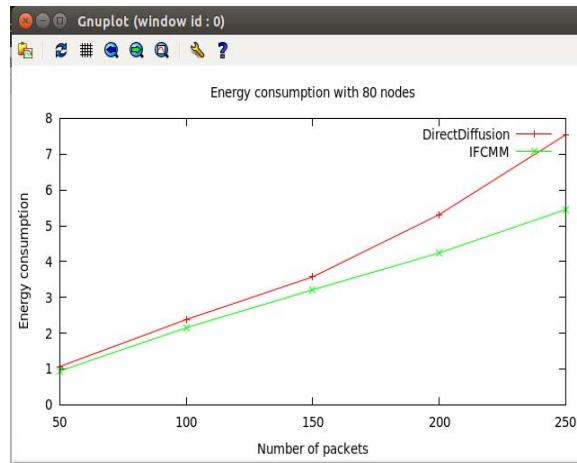


Figure 4.5: Energy consumption with $N=80$ nodes

4.2 Parameter Value

The proposed protocol is simulated using NS2 simulator with the following parameters as shown in below table 1.

Table 1: simulation parameters Proposed algorithm is evaluated by considering the below parameters:

Area	200m * 500m
Number of nodes	[50-100]
Number of SU	150
Number of channels	[2, 18]
Sensing time per channel	5ms
Channel switching time	80μs

4.3 Dead CH

The results in below Figure 4.6 show the comparison of dead CHs round by round. Dead CH refers to the nodes which are dead due to complete energy depletion or device failure. The results are compared with DEBR [12].

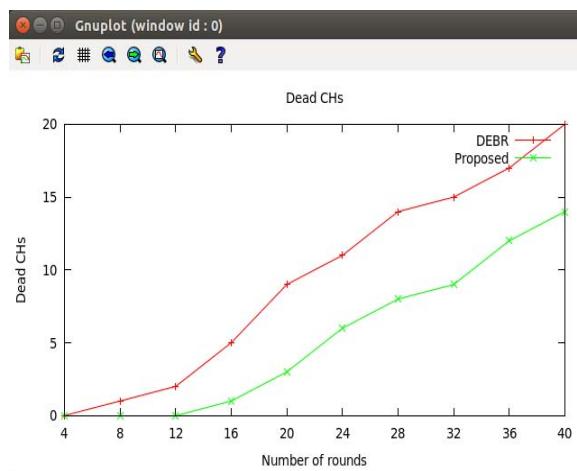


Figure 4.6: Dead Cluster Head vs. Number of nodes compared proposed with DEBR

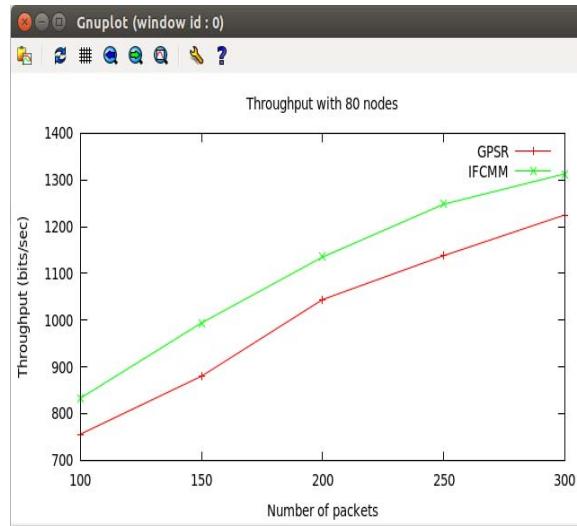


Figure 4.7: Throughput with 80 nodes

Figure 4.7 shows the throughput for total number of packets. In this graph we considered for 80 nodes. When throughput is compared with existing algorithm the proposed algorithm results in high throughput.

Figure 4.8 describes the delay for total 80 numbers of nodes. When compared to existing algorithm the delay is less in the proposed algorithm

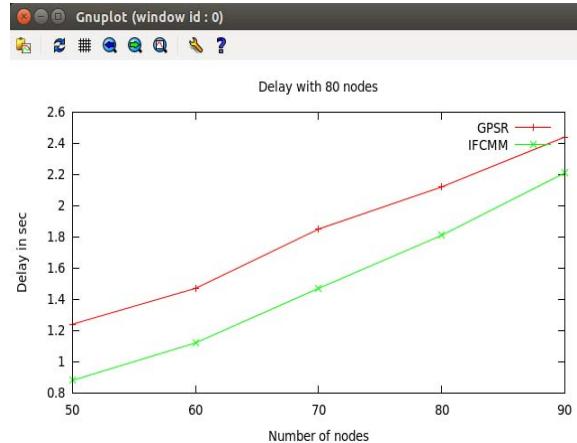


Figure 4.8: Delay with 80 nodes

V. CONCLUSION

This paper presents fault tolerant mechanism for multimedia. It considers the partition of the network into clusters. Energy conservation of the cluster head node is the main aim of this design. Cluster formation is based on residual energy of the CH. Proper selection of gateway nodes which act as replacement of CH node in case of failure will help in retaining the continuity of the network. Also, video streaming technique which operates in various spectrum condition is proposed, which takes higher bandwidth and low delay. The simulation result shows that the performance of the developed scheme is better than the earlier work.

REFERENCES

- [1] WANG, Beibei et LIU, KJ Ray. Advances in cognitive radio networks: A survey. *Selected Topics in Signal Processing, IEEE Journal of*, 2011, vol. 5, no 1, p. 5-23.
- [2] Akyildiz IF, Lee W-Y, and Chowdhury KR: CRAHNs: Cognitive radio ad hoc networks. *Ad Hoc Networks* 2009,7(5):810–836.
- [3] Jha SC, Rashid MM, Bhargava VK, Despins C: Medium access control in distributed cognitive radio networks. *Wireless Communications* 2011,18(4):41–51.
- [4] Jha SC, Phuyal U, Rashid MM, and Bhargava VK: Design of OMC-MAC: an opportunistic multi-channel MAC with QoS provisioning for distributed cognitive radio networks. *IEEE Trans. Wireless Commun* 2011,10(10):3414–3425.
- [5] Ehsan, Samina, and Bechir Hamdaoui. "A survey on energy-efficient routing techniques with QoS assurances for wireless multimedia sensor networks." *Communications Surveys & Tutorials, IEEE* 14.2 (2012): 265-278.
- [6] Priyanka Rawat, Kamal Deep Singh, Hakima Chaouchi, and Jean Marie Bonnin. "Wireless sensor networks: a survey on recent developments and potential synergies." *The Journal of Supercomputing* (2013), Springer
- [7] Bicen, A., Ozan, V., Cagri Gungor, and Ozgur B. Akan. "Delay-sensitive and multimedia communication in cognitive radio sensor networks." *Ad Hoc Networks* 10.5 (2012): 816-830.
- [8] Rehmani, Mubashir Husain, Stéphane Lohier, and Abderrezak Rachedi. "Channel bonding in cognitive radio wireless sensor networks." *Mobile and Wireless Networking (iCOST)*, 2012 International Conference on Selected Topics in. IEEE, 2012.
- [9] Liang, Zhongliang, and Dongmei Zhao. "Quality of service performance of a cognitive radio sensor network." *Communications (ICC), 2010 IEEE international conference on*. IEEE, 2010.
- [10] Liang, Zhongliang, et al. "Delay performance analysis for supporting real-time traffic in a cognitive radio sensor network." *Wireless Communications, IEEE Transactions on* 10.1 (2011): 325-335.
- [11] G. H. E. Fard, M. H. Yaghmaee, R. Monsefi, "An Adaptive Cross-Layer Multichannel QoS-MAC Protocol for Cluster Based Wireless Multimedia Sensor networks," Proc. of Ultra Modern Telecommunications & Workshops (ICUMT'09),pp.1-6, 12-14 Oct. 2009.
- [12] Y.-F. Wen, W. Liao, "On QoS Routing in Wireless Ad-Hoc Cognitive Radio Networks," Proc. of 71st IEEE Vehicular Technology Conference (VTC'10), pp.1-5, 16-19 May 2010.
- [13] L. Zhou, X. Wang, W. To, G. Mutean, and B. Geller, "Distributed Scheduling Scheme for Video Streaming over Multi-Channel MultiRadio Multi-Hop Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 28, no.3, pp. 409-419, Apr. 2010.
- [14] K.R. Chowdhury, M.D. Felice, "SEARCH: A routing protocol for mobile cognitive radio ad-hoc networks," *Computer Communications*, vol. 32 pp. 1983-1997, 2009.
- [15] Digham, F.F.; Alouini, M.S.; Simon, M.K. On the energy detection of unknown signals over fading channels. *IEEE Trans. Commun.* 2007, 55, 21–24.
- [16] KUMAR, Dilip, ASERI, Trilok C., et PATEL, R. B. EECDA: energy efficient clustering and data aggregation protocol for heterogeneous wireless sensor networks. *International Journal of Computers Communications & Control*, 2011, vol.6, no 1, p. 113-124.
- [17] T. Wiegand, and G. J. Sullivan, "The H.264/AVC Video Coding Standard," *IEEE Signal Proc. Mag.*, pp. 148-153 Mar2007.
- [18] SHAH, G., ALAGOZ, Fatih, FADEL, Etimad, et al. A Spectrum-Aware Clustering for Efficient Multimedia Routing in Cognitive Radio Sensor Networks. 2014.
- [19] Akyildiz IF, Su W, Sankara subramaniam Y. Wireless sensor networks: a survey. *Computer Network* 2002;38:393–422.
- [20] Abbasi AH, Younis M. A survey on clustering algorithms for wireless sensor networks. *Computer Commun* 2007;30:2826–41.
- [21] Akkaya K, Younis M. A survey on routing protocols for wireless sensor networks. *Ad Hoc Network* 2005;3:325–49.
- [22] Kuila P, Gupta SK, Jana PK. A novel evolutionary approach for load balanced clustering problem for wireless sensor networks. *Swarm Evol Comput*2013;12:48–56.
- [23] Bari A et al. Design of fault tolerant wireless sensor networks satisfying survivability and lifetime requirements. *Computer Commun* 2012;35(3):320–33.
- [24] Low CP et al. Efficient load-balanced clustering algorithms for wireless sensor networks. *Computer Commun* 2008;31(4):750–9.
- [25] Kuila P, Jana PK. Improved load balanced clustering algorithm for wireless sensor networks. In: *Proc. of int. conf. ADCONS 2011, LNCS* 7135; 2012. p.399–404.
- [26] Kuila P, Jana PK. Approximation schemes for load balanced clustering in wireless sensor networks. *J Supercomput* 2014;68(1):87–105.

- [27] Gupta G, Younis M. Fault-tolerant clustering of wireless sensor networks. Proc Int Conf IEEE WCNC 2003;3:1579–84.
- [28] Kuila P, Jana PK. Energy efficient load-balanced clustering algorithm for wireless Sensor Network. Proc Techno 2012;
- [29] Chiang SS et al. A minimum hop routing protocol for home security systems using wireless sensor networks. IEEE Trans Consum Electron 2007;53(4):1483–9
- [30] Ok Chang-So et al. Distributed energy balanced routing for wireless sensor networks. Computer Ind Eng 2009;57:125–35.
- [31] Heinzelman W, Chandrakasan A, Balakrishnan H. Application specific protocol architecture for wireless micro sensor networks. IEEE Trans Wireless Communication 2002;1(4):660–70.
- [32] Tyagi S, Kumar N. A systematic review on clustering and routing techniques based upon LEACH protocol for wireless sensor networks. J Network Computer Applications 2013;36:623–45.
- [33] Kuila P, Jana PK. An energy balanced distributed clustering and routing algorithm for wireless sensor networks. In: Proc. OF INT. CONF. PDGC 2012, IEEE Xplore; 2012. p. 220–225.
- [34]. Gurses, E., & Akan, O. B. (2005). Multimedia communication wireless sensor networks. *Annals of Telecommunications*, 60(7–8), 799–827.
- [35]. Akyildiz, I. F., Melodia, T., & Chowdhury, K. R. (2007). A survey on wireless multimedia sensor networks. *Computer Networks*, 51(4), 921–960.
- [36]. Mira, S., Reisslein, M., & Xue, G. (2008). A survey of multimedia streaming in wireless sensor networks. *IEEE Communications Surveys and Tutorials*, 10(4), 18–39. doi:10.1109/SURV.2008.080404.
- [37]. He, Z., & Wu, D. (2006). Resource allocation and performance analysis of wireless video sensors. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(5), 590–599.
- [38]. Karp, B., & Kung, H. T. (2000). GPSR: greedy perimeter stateless routing for wireless networks. In *Proceedings of the annual international conference on mobile computing and networking* (Mobi-Com 2000), Boston, USA, August.
- [39]. Kuhn, F., Wattenhofer, R., & Zollinger, A. (2003). Worst-case optimal and average-case efficient geometric ad-hoc routing. In *Proceedings of the 4th ACM international symposium on mobile ad hoc networking and computing* (MobiHoc 2003), Annapolis, MD, USA, June, 2003.
- [40] Kuhn, F., Wattenhofer, R., Zhang, Y., & Zollinger, A. (2003). Geometricad -hoc routing: of theory and practice. In *Proceedings of the 22nd ACM international symposium on the principles of distributed computing* (PODC 2003), Boston, Massachusetts, USA, July 13–16, 2003.
- [41] Leong, B., Mitra, S., & Liskov, B. (2005). Path vector face routing: geographic routing with local face information. In *Proceedings of the 13th IEEE international conference on network protocols* (ICNP 2005), Boston, Massachusetts, USA, November 6–9, 2005.
- [42] Gabriel, K., & Social, R. (1969). A new statistical approach to geographic variation analysis. *Systematic Zoology*, 18, 259–278.
- [43] Toussaint, G. T. (1980). The relative neighborhood graph of a finite planar set. *Pattern Recognition*, 12, 261–268.
- [44] Frey, H., & Stojmenovic, I. (2006). On delivery guarantees of face and combined greedy-face routing in ad hoc and sensor networks. In *Proceedings of the international conference on mobile computing and networking* (Mobi Com 2006), Los Angeles, USA, September, 2006.
- [45] Seada, K., Helmy, A., & Govindan, R. (2007). Modeling and analyzing the correctness of geographic face routing under realistic conditions. *Ad Hoc Networks*, 855–871. doi:10.1016/j.adhoc.2007.02.008.
- [46] Fang, Q., Gao, J., & Guibas, L. J. (2004). Locating and bypassing routing holes in sensor networks. In *Proceedings of the 23rd conference of the IEEE communications society* (INFOCOM 2004), Hong Kong, China, March, 2004.
- [47] Jia, W., Wang, T., Wang, G., & Guo, M. (2007). Hole avoiding in advance routing in wireless sensor networks. In *Proceedings of the IEEE wireless communication & networking conference* (WCNC2007), USA, March, 2007.
- [48] Yu, F., Lee, E., Choi, Y., Park, S., Lee, D., Tian, Y., & Kim, S. (2007). A modeling for hole problem in wireless sensor networks. In *Proceedings of the international wireless communications and mobile computing conference* (IWCMC 2007), Honolulu, Hawaii, USA, August, 2007.
- [49] Tsai, J., & Moors, T. (2006). A review of multipath routing protocols: from wireless ad hoc to mesh networks. In *Proceedings of A CoRN early career researcher workshop on wireless Multihop networking*, Sydney, July 17–18, 2006.
- [50] Johnson, D. B., & Maltz, D. A. (1996). Dynamic source routing in ad hoc wireless networks. In K. Imielinski (Ed.), *Mobile computing*. Dordrecht: Lower Academic.
- [51] Perkins, C. (2003). *Ad hoc on-demand distance vector (AODV)routing*. RFC 3561.

Constructions of Design Elements from Software Requirement Specifications (SRS)

Syed Naimatullah Hussain, CIT, Taif University, Taif, KSA
Email:s.naimatullah@tu.edu.sa

ABSTRACT

The software project development normally starts with customer's requirements. The customers will collect the requirements from the end users of the organization, and prepares a document called the SRS. Unified modeling language (UML) designed to provide a standard way to visualize the design of an information system. Apart from other internal lacunae of (UML) itself, the main reasons for low success rate are due to the non-availability of a correct and complete methodology for abstraction of the view elements for the design of an information system. This paper abstracts the view elements from Software Requirement Specification (SRS) for any information System in the form of Classes, Object of the Classes, Actors its interfaces their characterizing attributes, and these abstractions are further refined using good database design principles and the use of data flow diagrams (DFDs).

Keywords: Classes, Attributes, Interfaces

1. INTRODUCTION:

The software development project normally starts with SRS of the customers [1, 2]. The customers are in general, strategic management people of the organization who work in classical ambience. So the requirements of the expected system reflect their processing mindset. These requirements are influenced by either the data oriented approach or the procedure oriented approach as with the available information of the organization. Presently, since these are not natural ways of processing the information system, these will not serve the development process effectively. Now a day, people feel the object-oriented paradigm is more towards naturalness and will survive for long time. So it is required to transform the requirements into object-oriented paradigm(Classes, Object of Classes, Actors and its interfaces) and then proceed for the development. The main objective of this paper is to develop a system of software tools, that takes SRS and then transform it into Actors, Classes and Objects and its attributes.. There may also need to limit our ambition, as some of the sub processes may not be automated due to English as Natural Language. In such case, there is a need to provide guidelines for each of these sub processes to minimize the human dependency. the aim to develop a sequence of semi-automated software tools with embedded guidelines for inevitable subtasks at some stages then, the set of guidelines may give the scope to develop software agents to take up the role of semi automated processes.

Few researchers[3] have suggested some techniques for identification of object class hierarchies in procedural object oriented, but in our methodology we are using a perfect balance of

procedural as well data oriented paradigm i.e nothing but object oriented paradigm which more towards naturalness for certain stages of the design of object classes. [4, 9] in this authors are abstracting only object class and its attributes based on parsed use case description not all the design components from the SRS. Author's[8,10] this paper helps in the designing only the class and subclass structures but fails to identify the object class structure i.e. object class name, attributes, actors and its interfaces. Although, these give good guidelines to the design, the authors could not derive any concrete procedure and/or guidelines to the design in its totality. To develop a methodology that identifies Classes and Objects of the Classes, Actors and its attributes, , their characterizing attributes. This semi automatic methodology comprises of a sequence of steps like feasibility analysis, for object structure identification, resolution of synonyms & homonyms issues, regrouping of attributes of entities & functionalities through the design of data flow diagrams and elimination of imbalance data & procedure selection along with authentication of correctness & completeness of the abstractions at each stage. There is manual intervention at few stages is necessitated because of the need for human intelligence in these steps. Even for these manual intervention steps, attempt is made to provide clear-cut guidelines to streamline the design process.

2. Methodology

The algorithm presumes that, based on SRS from customer's this SRS is already available with the developer. Who can seek needy information from client's team of users (CTU). This paper addresses a sequence of semiautomatic methodology.

2.1 Requirement gathering.

As per the SRS, the detailed requirement is gathered from CTU. The input of any information system is SRS, output is Classes, Objects of the Classes, Actors and its attributes, If CTU is very large, then similar information is collected through the response to the questionnaire from the client.

2.2 Authentication of correctness and completeness of the process

Now we have two sets of entities, attributes, interrelationships, business rules, work processes and business process. The developer need to establish the one-to-one and on-to correspondence separately between each pair of items of two sets.

2.3 Resolve synonyms/homonyms issue.

As English is a Natural Language, in a multi-user system, though each user assigns meaningful

names to attributes, the semantic flexibility in the use of English words leads synonymous words for the same attribute. The set of synonymous words of the same meaning forms a synonymy and each such synonymy is replaced by a generic name. Similarly, the use of context specific word leads to homonyms. The presence of each such homonym in attributes/entities/actors should be replaced by different names.

2.4 Model the business process through to measure the paradigms imbalance.

- Entities, which are modifiable within the system
- Entities which are non modifiable within the systems form actors.

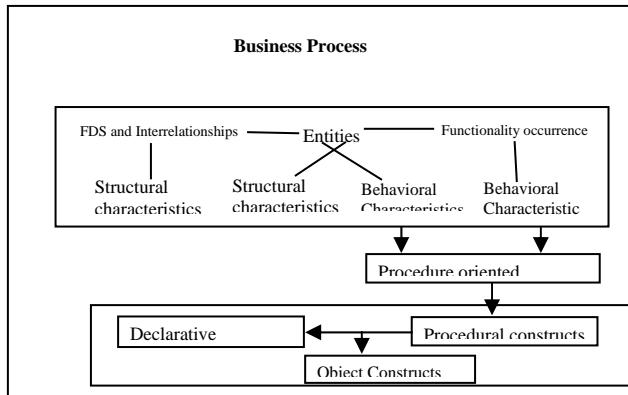


Fig. 1.1(Procedure Oriented)

2.5 Eliminate superfluity and redundancy in attributes/entities presence.

Study each attributes of each entity/actor in isolation with other entities/attributes for absolute necessity of their presence in the information system. This can be identified by the participation of the attribute in any of the functionalities. Discard the attributes that are not participating in the functionalities. If an attribute or group of attributes is present in two or more entities, form a separate entity with each such group. This participation can be tested through the establishment of one-to-one and on-to correspondence between attributes referenced in data flows of logical DFD (Fig 2.0)

2.6 Identify keys and design extended ER-diagram

Abstract the functional dependencies amongst attributes of each entity from the business rules of the information system. Identify the primary key and foreign keys for each entity/actor. If an attribute or group of attributes of an entity of data store is independent of the primary key, take it out and form separate entity. The entities and attributes are abstracted from the data dictionary and the interrelationships are abstracted from the business rules of the system.

2.7 Minimize the imbalance between the procedure and data oriented paradigms.

Figure-1.3 (Object Oriented)

[Fig-1.1 &1.2] Regroup the attributes of input data flows, based on their characterization of Person, place, thing, event

- The identified functionalities form processes
- The intersection of input/output attributes of a functionality with each entity attributes forms input/output dataflow from/to the concerned entity /actor to/form the process

This logical DFD (Fig-2.2) reveals the degree of imbalance in the paradigms opted while choosing structural and behavioral components.

Figures 2.51 and 2.52 depict the possible reasons tilt respectively towards Procedure oriented and data oriented paradigms

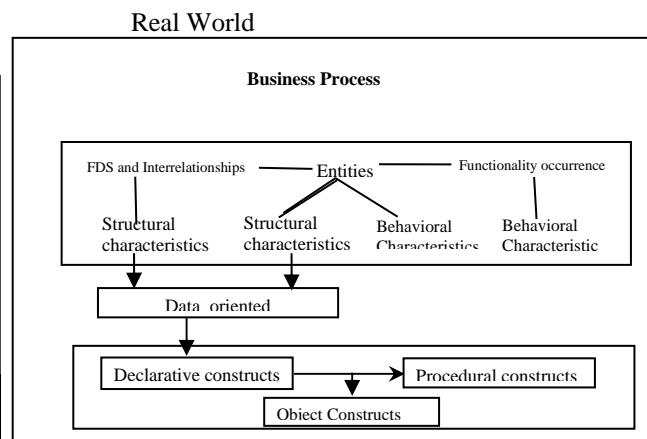
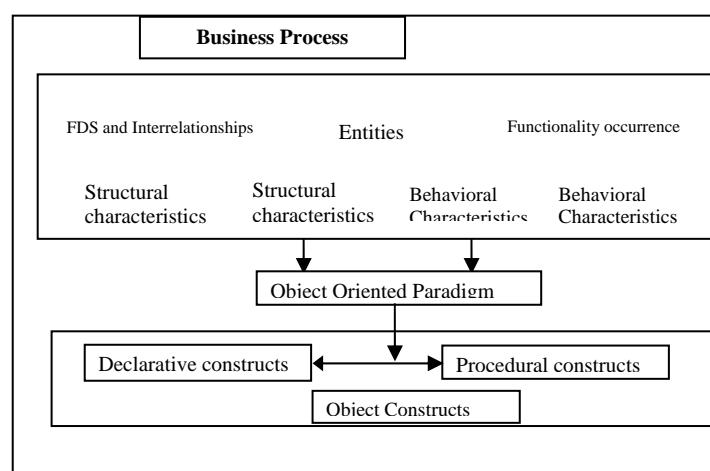


Fig.1.2 (Data Oriented)

or concept. This can be achieved through the grouping the attributes such that each attribute of a group establishes a functional dependence with one primary key. The input data flows may contain subset of attributes of a group so that each group of attributes may be in input data flows of one or more processes. Each such group forms a first cut object structure. [Fig-1.3] shows the perfect balance between data and procedure oriented i.e object oriented which is more towards naturalness.



2.8 Refine the abstracts by bringing the perfect balance between the paradigms.

[5.6.7]

Apply good database design principles to each first cut object structure [section 3.2].

Oriented

Since the normalization is a way of good database design, the refined group should be normalized at least up to Boyce coded normal form. Figure-2.2 indicates the constructs to bring perfect balance between procedure oriented and data oriented paradigms.

Figure-1.3 Object

2.9 Design the logical DFD with object structure

Fig-2.2 Refine the logical data flow diagram with each object structure group as data store, maintaining attributes of total input data flow to each process and redesigning input flows such that each flow emanates from first cut object structure.

2.10 Decompose functionality

If two or more data flows directed towards single process, study the possibility of decomposing the process, so that each decomposed process either receives data flow from single entity or additionally through parameter passing from other entities. Refine the logical data flow diagram into physical data flow diagram.

2.11 Software Requirement Specification[SRS]

The input to the any information system is the requirements that contain the business rules of the information system along with various applications. For example, SRS for Banking Application information system may contain some of the business rules as follows.

3.0 Theory

Automate the customer transaction in the banking system
There are two types of customer.

Deposit customer

- Each customer is provided with a unique account number
- Periodically credits/debits from his/her account.
- The customer earns interest on minimum balance of each month.
- There may be overdraft facility for some customer who has developed some kind of rapport with the bank.
- The customer can also transfer amount from his/her account to any account.

Loan Customer

- Loan customer initially applies for loan for specific purpose.
- The bank grants loan to customers after authenticating their repayments in installments.
- The bank charges interest either on floating-point rate or fixed-point rate.
- Interest rate for bank loan is generally much higher than interest rate for bank deposit
- Thus the bank earns profit by liaising between savings customer and loan customer
- Each branch manager needs to submit periodical summary report indicating the total number of transactions, the total amount deposited, the total interest paid to the customer, the total amount sanctioned, the interest earn by customer and thus the profit earn by the bank in the specific

3.1 Identification of Classes, attributes from SRS

Depositor (Acctno,D-name,D-address,Contact-no)
Transaction (Acctno,Pre-balance,Transtype,Transamt,New-balance,OD-facility)
Daccount (acctno, period, minibalance, int-rate, intAmt)
Transfer (acctno, minibalance,ch-amt,payee-acct,newbalance,payeebalance)
Borrower (acctno, loan-type, Amt-sanctioned, Amtavailed, newbalance, int-rate)
EMI (acctno, Balance, Rp-period, Principal, Int, EMI)
Baccount (acctno, period, maxbal, int-rate, intamt)
Report (period, no of depositor, amt deposited, int-paid, amt-sanctioned, no of borrowers, amt-disbursed, int-earned, profit)
Loan (Bacctno, Bname, Baddress, Bcontact no)

3.2. Synonyms and homonyms

These are Homonyms found in various entities like, Acctno, period, Int-rate, Int-amt, NewBalance.
Depositor (Acctno,D-name,D-address,Contact-no)
Transaction (Acctno,Pre-balance,Transtype,Transamt,New-balance,OD-facility)
Daccount (Dacctno, Dperiod, minibalance, Dint-rate, DintAmt)
Transfer (Dacctno, minibalance,ch-amt,payee-acct,Dnewbalance,payeebalance)
Borrower (Bacctno, loan-type, Amt-sanctioned, Amtavailed, Bnewbalance, Bint-rate)
EMI (Bacctno, Balance, Rp-period, Principal, Int, EMI)
Baccount (Bacctno, Bperiod, maxbal, Bint-rate, Bintamt)
Report (period, no of depositor, amt deposited, int-paid, amt-sanctioned, no of borrowers, amt-disbursed, int-earned, profit)
Loan (Bacctno, Bname, Baddress, Bcontact no)
I could not fine any of synonyms in this application

4. DATA FLOW DIAGRAM (DFD) (Context Level Diagram and Logical DFD refer Page no.4, 5)

Results:

Actors: Customer, Depositor

Attributes of Actor: report, transaction, EMI, Bank database

Classes: Sanction Loan, Transaction Amount, Transfer Amount, Compute EMI, Prepare Report, Assign Account No.

Objects of Classes: Acctno,D-name,D-address,Contact-no,Acctno,Pre-balance,Transtype,Transamt,New-balance,OD-facility) Dacctno, minibalance,ch-amt,payee-acct,Dnewbalance,payeebalance
(Bacctno, Balance, Rp-period, Principal, Int, EMI)
period, no of depositor, amt deposited, int-paid, amt-sanctioned, no of borrowers, amt-disbursed, int-earned, profit, Bacctno, Bname, Baddress, Bcontact no.

Discussion:

Unified Modeling Language [UML][11,12] has become a de facto standard for the design and implementation of object oriented information system. Though the UML [13] is a boom for the design of business processes in an object oriented paradigm, the success rate of projects involving the design using UML is very slow.

The SRS is the abstraction of number of end user's requirements. Because of the environment in which the individual is working, the end user may use different words for the same meaning. Thus, the SRS contains synonymous words. In addition, the use of contexts specific words compels the end user to use the same word for different meanings. Thus, SRS also contains homonymous. The non-resolution of synonymous issue results in the redundancy model elements whereas; non-resolution homonym issue results in violation of uniqueness of the model elements. Thus any information system design needs to resolve the synonym and homonym issues. Thus, the success rate of the information system developed using such error prone approaches swings. There is an urgent need to isolate the dependency of software development success rate on human skills and to elevate it to the height of the automatic process.

5. CONCLUSION:

UML is a graphical language for visualizing, specifying, constructing, and documenting the artifacts of a software-information system. But suffer from the correctness and completeness aspect. Developed methodology helps in designing, design components (Classes, Objects and its Attributes) in a semiautomatic way incorporating correctness and completeness in each of its stages. This methodology minimizes human interventions that too giving clear-cut guidelines. In future these, intervention also can be automated.

6. REFERENCES:

- [1] Pankaj Jalote (2005). An Integrated approach to Software engineering, 3rd edition springer's publications, India.
- [2] Roger S Pressman (2005). Software Engineering, 6th edition McGraw-Hill companies international edition, India
- [3] Mosa Emhamed Elbendak(2005) "Framework for using a Natural Language Approach to Object Identification Framework for using a Natural Language" <http://www.aclweb.org/anthology/R09-2005>.
- [4] Muhammed Usman, Stephane Ducane and Marianne Huchard (IEEE-2008) 15th working conference on reverse engineering "Reconsidering classes in procedural object oriented code" page 257- 266.
- [5] Jonathan & Charles J Hannon (IEEE-2009) Eighth Mexican International Conference on Current Trends in Computer Science. "An algorithm for identifying Authors using synonyms" page 99 – 104.
- [6] Alen Lovrencic and Tonimir Kisasondi (IEEE-2007) 11th International Conference on Intelligent Engineering System "Modelling Functional Dependencies in Databases using mathematical logic" page 307 – 312.
- [7] Steffen Rudle and LarsSchmidt-thieme (IEEE-2006) Sixth International Conference in Data Mining "Object Identification with constraints" page 1 -7.

- [8] Sukhamay Kundu & Migel (IEEE-2005) 29th International Conference Software and Application Conference "A Formal approach to Designing a class subclass structure using a partial order on the functions" page 1 -8.
- [9] Mosa Emhamed Elbendak "Framework for using a Natural Language Approach to Object Identification Framework for using a Natural Language" Student Research Workshop, RANLP 2009 - Borovets, Bulgaria, pages 23–28
- [10] .S. Anandha Mala and Dr.G.V.Uma "Object Oriented Visualization of Natural Language Requirement Specification and NFR Preference Elicitation. IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.8, August 2006.
- [11] Lilac A. E. Al-Safadi "Natural Language Processing for Conceptual Modeling" A semi-automatic approach for designing databases. International Journal of Digital Content Technology and its Applications Volume 3, Number 3, September 2009
- [12] Sunguk Lee "UML for Database System and computer application. Interantional Journal of Database System and Computer Application. Volume 5, number 3, 2012
- [13] K.P Jayant, Garg,Kumar, Rana. "An approach of Software Design Testing Based on UML diagram. International Journal of Advanced Research in Computer Science and Software Engineering. Volume-4, Issue-2, February 2014
- [14] Tiwari,Alpika,Dubey "Merging of Dataflow Diagrams with Unified Modeling Language" Internation Journal of Scientific and Research Publication Volume-2,issue-8, August-2012
- [15] Rosziati,Yen "A Formao Model for Data Flow Diagrams Rules" APRN Jounal of System and Software, volume-1 No.2, May-2011

Context Level Diagram [14]:

Fig-2.1 Depicts that the interaction between the banking information systems with the external agents which acts as data sources and data sinks.

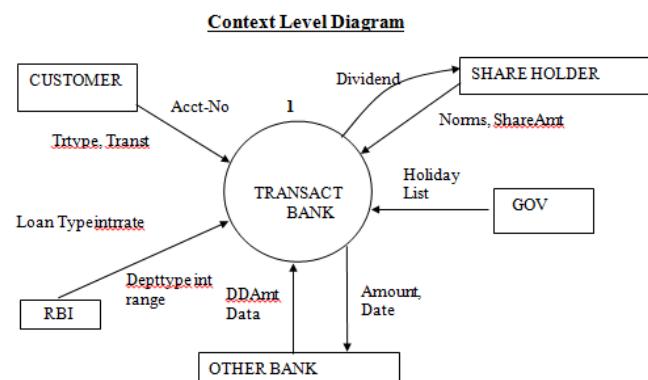


Figure-2.1 Context Level Diagram

Figure-2.2 logical DFD [15] captures the data flows that are necessary for a banking information system to operate. It describes the processes that are undertaken, the data required and produced by each process.

#

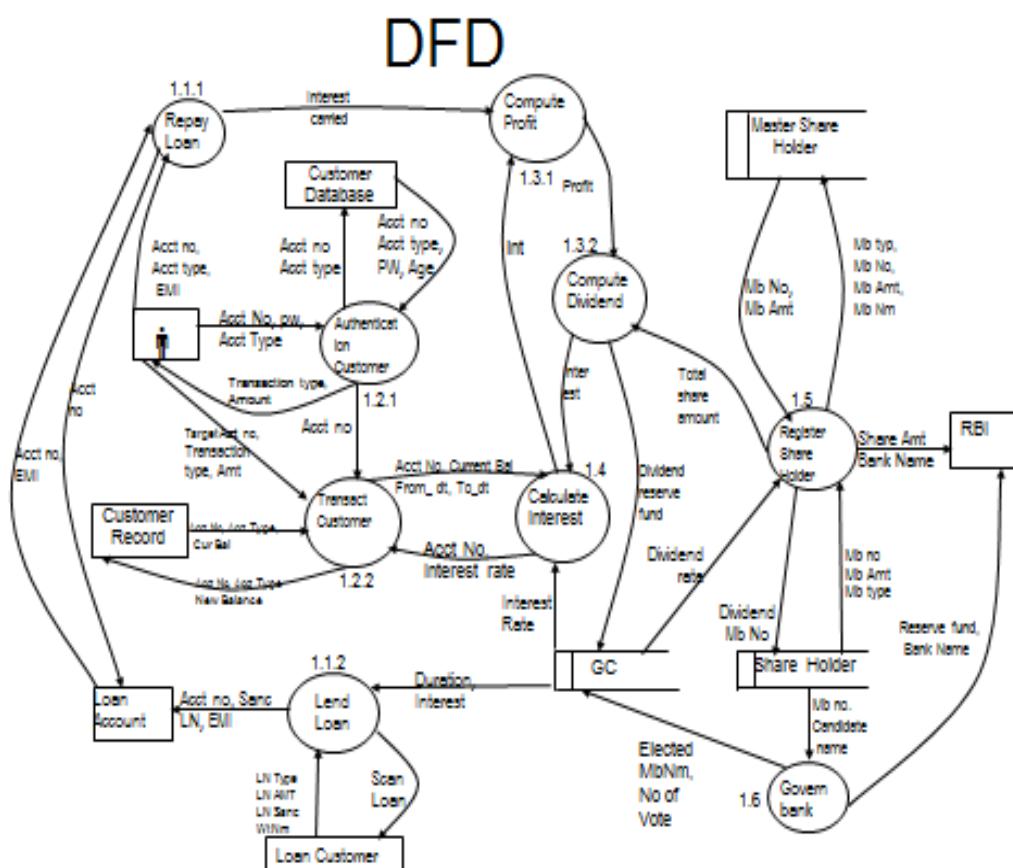


Figure-2.2 Logical DFD

An Efficient Way of Medical Image Encryption using Cat Map and Chaotic Logistic Function

Ranu Gupta ^{1*}, Rahul Pachauri ², Ashutosh K Singh ³

^{1,2}Jaypee University of Engineering and Technology, Raghogarh, Guna (M.P.) India, 473226

³Thapar Institute of Engineering and Technology University, Patiala, 147004, India

*Corresponding author, E-mail: ranugupta02@yahoo.com

pachauri.123@gmail.com⁽²⁾, aksingh@thapar.edu⁽³⁾

Abstract: Encryption is used to securely transmit data in open networks. The proposed image encryption method uses Arnold cat map and one dimensional chaotic map. The pixel shuffling is done using cat map. Then the encryption is done by using 128 bit long external secret key. The initial condition is calculated using secret key. Two matrices of size of the image are formed using chaotic logistic function. Finally the image is encrypted by performing XOR operation with the two matrices formed by the chaotic function and the shuffled image.

I INTRODUCTION

In today's Hi-Tech world of innovations in the field of internet, medical imaging systems, military message communication, it is needed to transmit the images through network confidentially. For reliable and secure image transmission, we heavily rely on image cryptography which is the base of security in digital communication. Many image encryption methods are available to protect the content of digital images but, few of them are at par with the expected goals i.e. speed, reliability and security. Since the images play a vital role in the field of medical treatment of the patient. These medical images need to be transmitted through public network in order to consult the doctors. Therefore security plays a vital role while transmitting the images. While the general images are important in everyday life. The traditional encryption schemes like simple-Data Encryption Standard (DES), triple- DES, Rivest Shamir Adleman (RSA), International Data Encryption Algorithm (IDEA) and Advanced Encryption Standard (AES) do not fit for modern image transmission requirement. Many researchers have tried to innovate better solutions for image encryptions. In particular, application of chaos theory in multimedia encryption is one of the important research directions. The field of chaotic cryptography has undergone tremendous growth over the past few decades. The primary motivation of employing chaotic systems is its simplicity in form and complexity in dynamic are not appropriate to make cryptosystems for large digital data. For encrypting the digital images, plenty of encryption schemes have been proposed [1-15]. Scanning procedure for the encryption and simultaneously applying compression to the image [2] was proposed. In [6], a symmetric encryption scheme based on 2D chaotic map is proposed. A two or higher dimensional discretized chaotic maps is adopted for pixel permutation together with 1D map for diffusion. The superiorities of such kind of chaos-based approaches are mainly relatively large block size and a high encryption rate. The analysis of nonlinear chaotic algorithm (NCA) map and a no. of attacks were proposed in [7]. Image encryption scheme based on 3D baker with dynamical compound chaotic sequence cipher generator was proposed in [8]. Divide dynamic block of 3D baker map by using compound chaotic map, and compare with 2D baker map. The 3D baker scheme is 2-3 time faster of 2D baker map. The proposed scheme used in real-time secured image transmission.

This work combines the confusion/diffusion in single unit for image encryption. Permutation and diffusion are two separate and iterative stages, and they both require scanning the image in order to obtain the pixel values in fastest way. The paper is organized as follows:

Section-2 and 3 gives a short overview of chaotic logistic function and the proposed method respectively. Section 4 evaluates the proposed method with the performance parameters. Section-5 gives the conclusion.

II CHAOTIC LOGISTIC FUNCTION

The chaotic logistic map is random and simple to implement and therefore cryptographers are diverted to this approach for image encryption. The one dimensional chaotic function [1] is expressed as:

$$X_{i+1} = aX_i(1 - X_i) \quad (1)$$

where X_i takes values in the range from [0-1]. It is the simplest method in chaos. Assuming initial condition to be $X_0 = 0.4$, for different value of a , the logistic equation is evaluated through simulation and shown in Fig. 1. It can be concluded that when a is in the interval [3.5-4], it is highly chaotic.

The characteristics of chaotic systems are [2]:

- i It is deterministic and follows some mathematical equation.
- ii But it appears to be highly random.
- iii It is unpredictable and follows non linear relation.
- iv It is very sensitive to initial condition i.e with a slight change in the initial condition there is a vast change in the succeeding values.

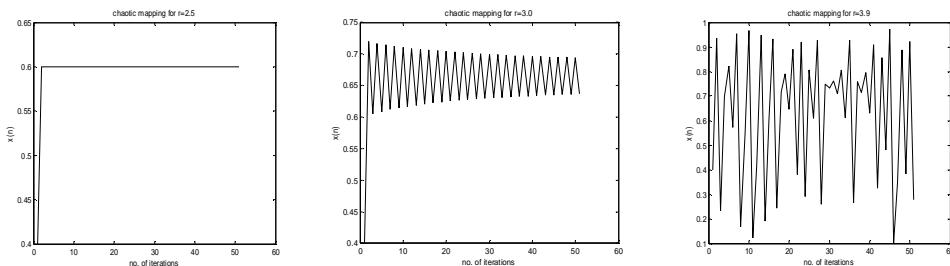


Figure 1. Iteration property

III PROPOSED METHOD

In the proposed image encryption method, the image is permuted by Arnold cat map. Thereafter, the external secret key of 16 ASCII characters is entered. The flow diagram of the proposed method is shown in Fig. 2. Steps involved in the proposed method for the encryption of the image are as follows:

- (a) The image is shuffled by using Arnold cat map.

$$\begin{aligned} x' &= (x + ay) \bmod(N) \\ y' &= (bx + (ab + 1)y) \bmod(N) \end{aligned} \quad (2)$$

Where a, b are controlling parameters which are positive integers and x', y' is the new pixel position of the original image pixel of position x, y respectively. It is applied once in the image.

- (b) A secret key of 16 ASCII characters (128-bit long) is used in the proposed method. The secret key can be represented as:

$$K_i = K_1 K_2 K_3 \dots A_{16} \text{ (in ASCII)} \quad (i=0-16) \quad (3)$$

(c) The initial condition (X_0) for the chaotic map is calculated as follows:

$$T = (\sum_{i=1}^{16} S_i * (K_i)) \bmod 256 \quad (4)$$

$$X_0 = T - \lfloor T \rfloor \quad (5)$$

$$\text{where, } S_i = 0.123 + S_{i-1} \quad (6)$$

and K_i , $\lfloor \cdot \rfloor$, are the decimal equivalent of the keys, the floor function respectively. The initial value of S_i is taken to be zero.

(d) A chaotic logistic map as in equation (1) is used in the proposed method to generate two matrices of the same as image for image encryption. The initial values calculated are ranging from 0-1, therefore these values are converted into whole numbers < 256 .

(e) Then one by one the consecutive byte is read from the shuffled image. The encryption is done by XORing the corresponding values of the three matrices. Two matrices X_i , Y_i generated from logistic function and third is the image matrix P_i .

$$C_i = P_i \oplus X_i \oplus Y_i \quad (7)$$

(f). In case of decryption process all the steps from (a) to (e) are same but decryption is done in the reverse order.

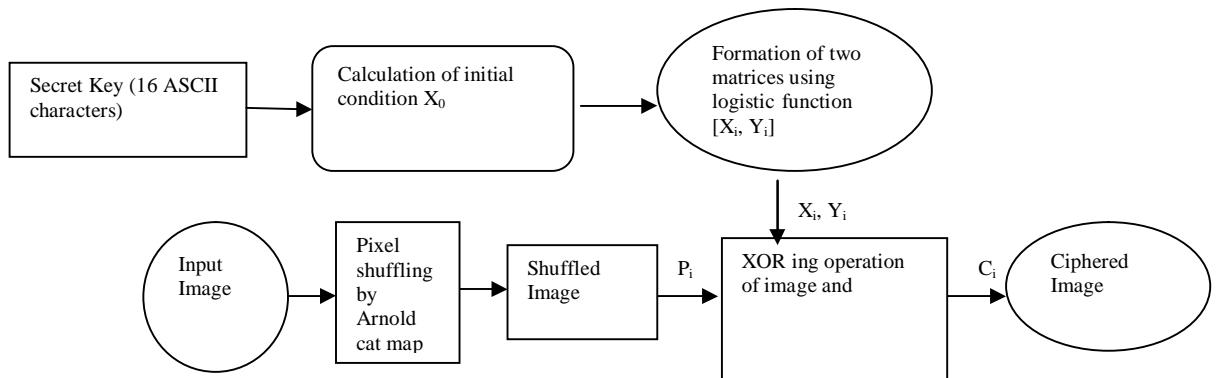
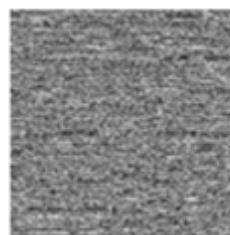


Figure 2. Block Diagram of Proposed Method



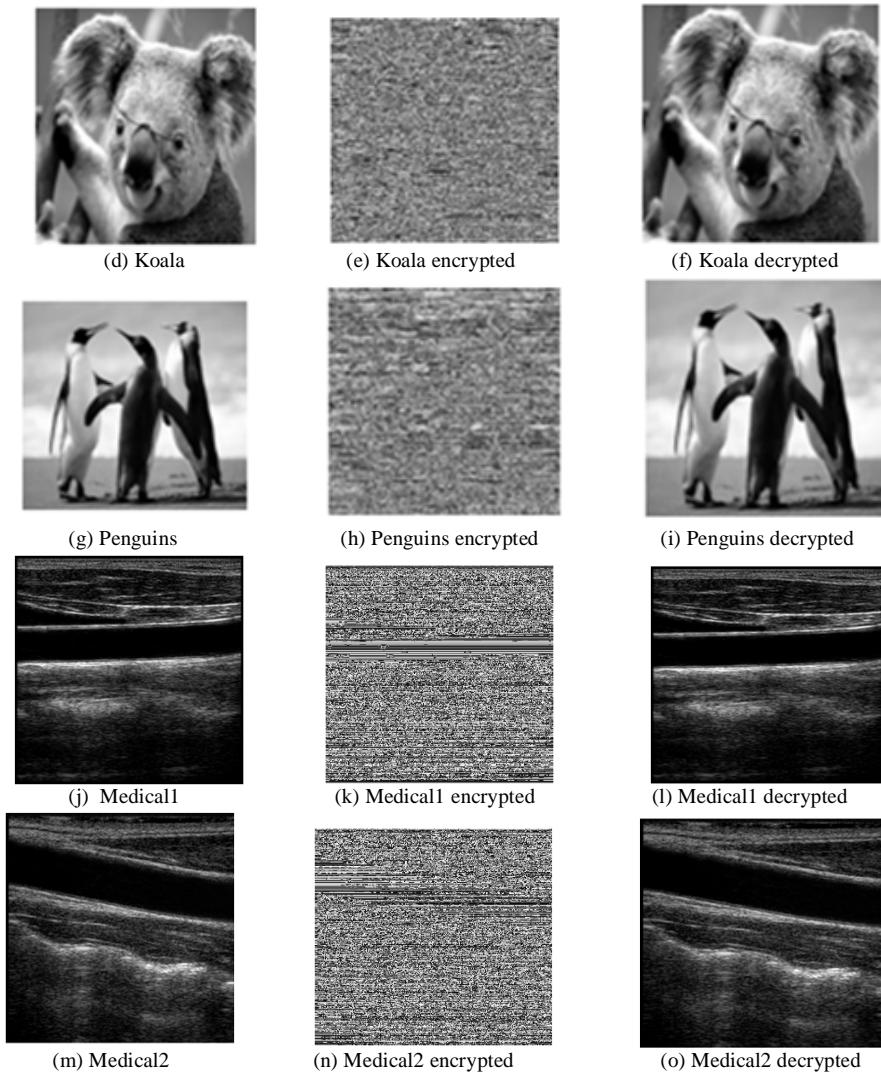


Figure 3. Original and encrypted images with the proposed method

As shown in Fig. 3 five images (a), (d), (g), (j), and (m) are encrypted by the proposed method whereas Fig. 3 (b), (e), (h), (k), and (n) shows the encrypted images. The encrypted images show that it is quite difficult to trace the original information, making the proposed method more efficient. While Fig. 3 (c), (f), (i), (l), and (o) shows its corresponding decrypted images.

IV[~] PERFORMANCE ANALYSES

The various performance analyses are done of the proposed method in the following sections:-

A Histogram Analysis

The histograms of various encrypted and original images is shown in Fig. 4 (b), (e), (h), (k), and (n) and Fig. 4 (a), (d), (g), (j), and (m) respectively.

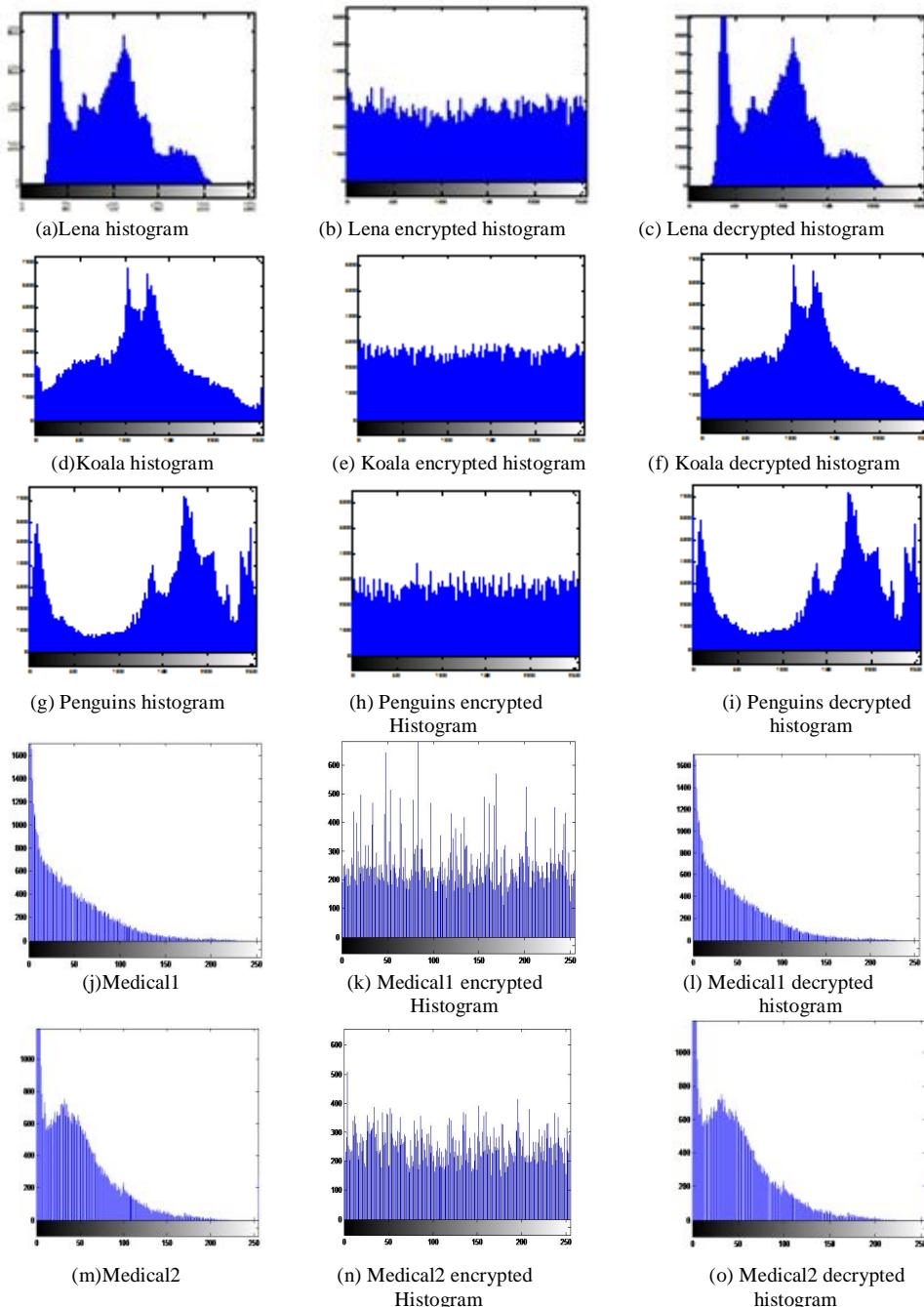


Figure 4. Histogram Analyses of Original & Encrypted Images

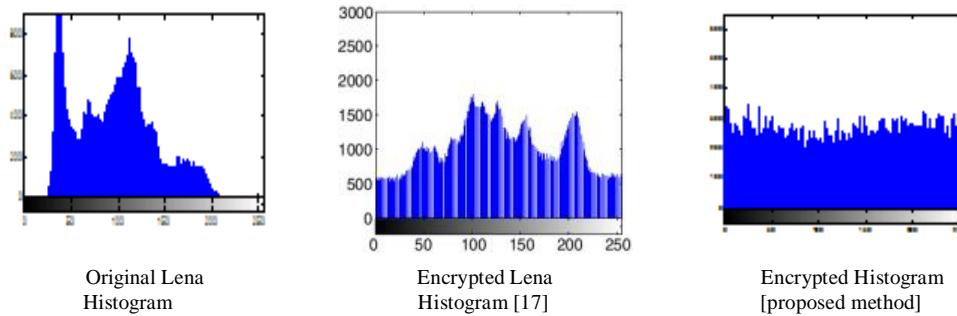


Figure 5. Comparison of histogram with the proposed method

B Correlation Coefficient Analysis

In addition to the histogram analysis, we have also analyzed the correlation between two vertically adjacent pixels, two horizontally adjacent pixels and two diagonally adjacent pixels in plain-image/cipher-image respectively. Following formula is used for calculation [4].

$$C_r = \frac{N \sum_{j=1}^N (x_j \times y_j) - \sum_{j=1}^N x_j \times \sum_{j=1}^N y_j}{\sqrt{\left(N \sum_{j=1}^N x_j^2 - (\sum_{j=1}^N x_j)^2 \right) \times \left(N \sum_{j=1}^N y_j^2 - (\sum_{j=1}^N y_j)^2 \right)}} \quad (8)$$

where x and y are the two neighboring pixels and N is number of pixels in the image. Table I and II shows the correlation coefficients of the original as well as encrypted images respectively. The images are encrypted using the secret key “12ghUO3456HJKLjx”. Table II shows that encrypted image are weakly correlated. The pictorial representation of coefficient of correlation of Lena image and the encrypted image with the proposed method is shown in Fig. 6. The comparison of coefficient of correlation of the proposed method with other method is shown in Table III.

Table I Correlation coefficients for the neighboring pixels in the original images

S.No.	Images	Size	Vertical	Horizontal	Diagonal
1.	Lena	256x256	0.9414	0.9204	0.9412
2.	Koala	683x512	0.9553	0.9470	0.9553
3.	Penguins	1024x768	0.9785	0.9576	0.9785
4.	Medical1	348x412	0.8801	0.9836	0.8801
5.	Medical2	389x367	0.8278	0.9194	0.8278
6.	Medical3	400x307	0.9275	0.9718	0.9275

Table II Correlation coefficients for neighboring pixels in the encrypted images

S.No.	Images	Size	vertical	horizontal	Diagonal
1.	Lena	256x256	-0.00035	0.00074	-0.00013
2.	Koala	683x512	0.00064	0.00066	-0.0006
3.	Penguins	1024x768	0.00062	0.00124	0.0035
4.	Medical1	348x412	0.00151	0.00121	0.0031
5.	Medical2	389x367	0.00085	0.00104	0.0041
6.	Medical3	400x307	0.00271	0.00144	0.0075

Table III Comparison of correlation coefficient of the proposed method (Lena)

Correlation Coefficient	[18]	[19]	[20]	[21]	[22]	[23]	Proposed method
Vertical	0.028	0.040	0.065	0.027	0.0024	0.0021	-0.0035
Horizontal	0.045	0.082	0.016	0.012	-0.0086	0.0046	0.0074
Diagonal	0.021	0.005	0.032	0.007	0.0402	0.0033	-0.0013

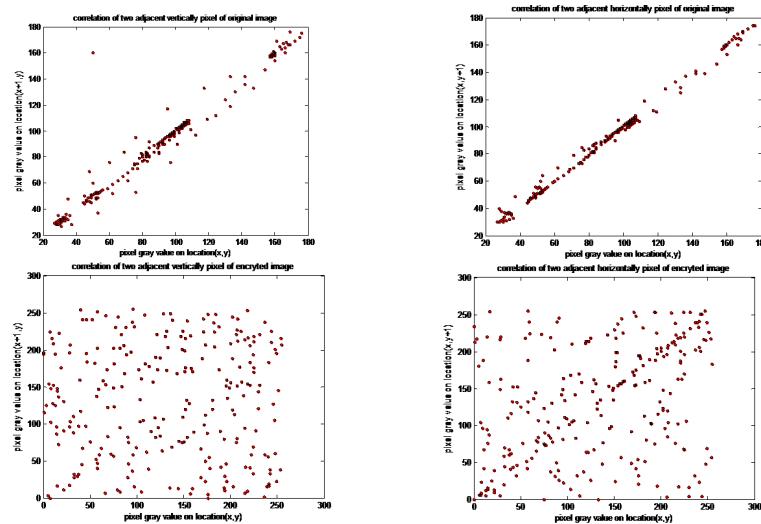


Figure 6. Horizontal and vertical Correlation coefficient of Lena original and encrypted image

C Sensitivity Analysis

Sensitivity analysis can be done on the secret key as well as on the image. To test the sensitivity of the proposed method the original image is encrypted by three different session keys, first by changing the MSB and second by LSB. The three session keys are ‘12ghUO3456HJKLjx’, ‘M2ghUO3456HJKLjx’, and ‘12ghUO3456HJKLj4’ respectively. The encrypted images and its correlation are shown in Fig. 7 and Table IV.

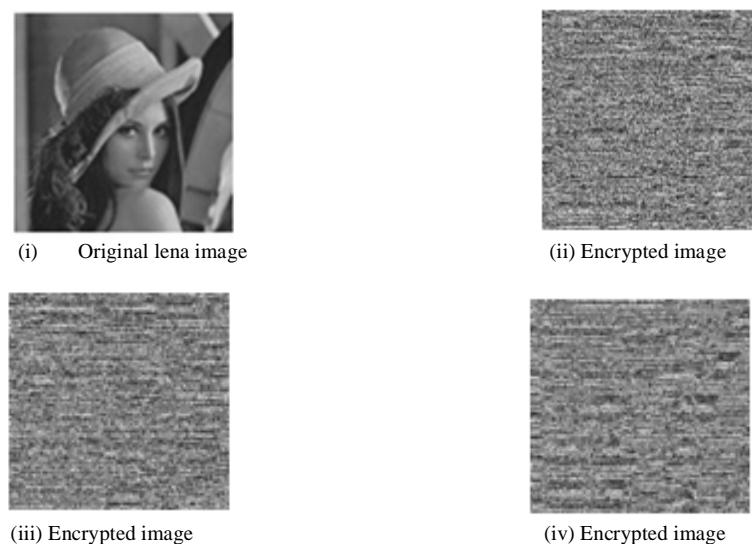


Figure 7. Sensitivity test I: Frame (i) Original image, (ii), (iii), (iv) Encrypted images

Table IV Correlation coefficients of the three encrypted images

S.No.	Images	Images	Correlation coefficient
1.	(ii)	(iii)	-0.00084
2.	(iii)	(iv)	-0.00075
3.	(iv)	(ii)	-0.00043

Table IV shows the correlation coefficients between the corresponding pixels of the three encrypted images (ii), (iii) and (iv) and indicate that the proposed method is very sensitive to even a slight change in the secret key.

D Number of Pixels Change Rate (NPCR)

The NPCR [10] is defined by the following equation:

$$NPCR = \sum_{i=1, j=1}^{x, y} \frac{D(i, j)}{x \times y} \times 100\% \quad (8)$$

If $C_1(i, j) = C(i, j)$ then $D(i, j) = 0$ else
 $D(i, j) = 1$

where x and y are the width and height of encrypted image. The NPCR for various images is calculated by the proposed encryption method and found to be above **99%**. Table V shows the comparison of the NPCR with the proposed method.

E Unified Average Change Intensity (UACI)

UACI can be defined as

$$UCAI = \sum_{i=j=1}^{w, h} \frac{|E1(i, j) - E2(i, j)|}{255wh} \times 100 \quad (9)$$

where w and h are the width and height of the image. It shows the average change in the intensities of the pixel. The comparison of UACI with other methods is shown in Table V.

Table V Comparison of NPCR and UACI criteria of proposed method and the other methods for Lena image

Methods	NPCR %	UACI %
[20]	NA	NA
[21] 2 nd round	25	8.50
[18] 1 st round	37	9
[26]	97.62	32.90
[19]	98.669	33.362
[25]	99.52	33.14
[23]	99.63	33.71
[27]	99.61	33.35
[24]	99.70	29.3
Proposed Method	99.9975	29.6477

G Time Analysis

The time is also calculated for the encryption and decryption of Lena image by the proposed method. The time analysis is done on Intel(R) Core (TM) 2 Duo CPU T5870 @2.00GHz with 3GB RAM computer. The coding is done on MATLAB 7.9.0(R2009b). Table VI shows the encryption/ decryption time taken by the proposed method.

Table VI Time taken in Encryption and Decryption by the proposed method

Image	Size	Encryption Time (secs)	Decryption Time (secs)
Lena	256x256	0.00222775	0.00223049

V CONCLUSION

In this work a technique is proposed which replaces the traditional preprocessing complex system and utilizes the basic operations like confusion, diffusion which provide better encryption using cat map and chaotic function. The proposed method is successfully and efficiently implemented to various images. The performance of various parameters shows that the proposed method is robust efficient, secured and fast. It can also be used in real time applications. Theoretical analyses and computer simulations on the basis of histogram analysis, correlation analysis, NPCR and UACI confirm that the new algorithm minimizes the possibility of brute force attack for decryption and fast for practical image encryption. Its future scope is that the method can be varied by increasing the key length and modifying the mathematical calculation for calculating the initial condition.

DISCLOSURE

There was no funding from any agencies either government or non-government.

ACKNOWLEDGMENTS

The general images were taken from the site <http://sipi.usc.edu/database/> which is freely available whereas medical images were taken from Government medical college, India.

REFERENCES

- [1] G. Chen, X.Y. Zhao, “A self-adaptive algorithm on image encryption”, International Journal Software, 16, 1987, pp. 1975–1982.
- [2] N. Bourbakis, C. Alexopoulos, “Picture data encryption using SCAN patterns”, Pattern Recognition, 25(6), 1992, pp. 567–581.
- [3] K.L. Chung, L.C. Chang, “Large encrypting binary images with higher security”, Pattern Recognition Letters, 19(5-6), 1998, pp. 461– 468.
- [4] J.C. Yen and J.I. Guo, “A new image encryption algorithm and its VLSI architecture”, IEEE Workshop on Signal Processing System, 3 1999, pp. 430–437.
- [5] H. Cheng, X.B. Li, “Partial encryption of compressed images and videos”, IEEE Transaction in Signal Processing, 48(8), 2000, pp. 2439– 2451.
- [6] J. Fridrich, “Symmetric ciphers based on two-dimensional chaotic maps”, International Journal of Bifurcation and Chaos, 8(6), 1998, pp. 1259–1284.
- [7] G. Alvarez, Shujun Li, “Crypt analyzing a nonlinear chaotic algorithm (NCA) for image encryption”, Commun Nonlinear Sci. Numer. Simulation, 14, 2009, pp. 3743– 3749.
- [8] M.S. Baptista, “Cryptography with chaos”, Physics Letter A, 240(1-2), 1999, pp. 50–54.
- [9] X. Liao, X. Li, J. Pen and G. Chen, “A digital secure image communication scheme based on the chaotic chebyshev map”, International Journal of Communication Systems, 17(5), 2004, pp. 437–445.
- [10] F. Han, X. Yu and S. Han, “Improved baker map for image encryption”, International Symposium on Systems & Control in Aerospace & Astronautics, 2, 2006, pp. 1273–1276.
- [11] S. Lian, J. Sun, Z. Wang, “A block cipher based on a suitable use of chaotic standard map”, Chaos Solitons & Fractals, 26(1), 2005, pp. 117–129.
- [12] B. He, F. Zhang, L. Luo, M. Du, Y. Wang, “An image encryption algorithm based on spatiotemporal chaos”, International Congress on Image and Signal Processing, 2009, pp. 1–5.
- [13] G. Chen, Y. Mao, C.K. Chui, “A symmetric image encryption schemes based on 3D chaotic cat maps”, Chaos Solitons and Fractals, 21(3) 2004, pp. 749–761.

- [14] Z.H. Guan, F. Huang, W. Guan, "Chaos-based image encryption algorithm", Physics Letters A, 346(1-3), 2005, pp. 153–157.
- [15] T. Gao, Z. Chen, "Image encryption based on a new total shuffling algorithm", Chaos Solitons & Fractals, 38(1), 2008, pp. 213–220.
- [16] Xiaojun Tong, MinggenCui, "Image encryption scheme based on 3D baker with dynamical compound chaotic sequence cipher generator", Signal Processing, 89, 2009, pp. 480–491.
- [17] G. Zhang, Q. Liu, "A novel image encryption method based on total shuffling scheme", Optics Communications, 284 (12), 2011, pp. 2775–2780.
- [18] Y. Mao, G. Chen, S. Lian, "A novel fast image encryption scheme based on 3D chaotic baker maps", International Journal of Bifurcation and Chaos, 14(10), 2004, pp. 3613–3624.
- [19] L. Zhang, X. Liao, X. Wang, "An image encryption approach based on chaotic maps", Chaos Solitons & Fractals, 24 (3), 2005, pp. 759–765.
- [20] H. Gao, Y. Zhang, S. Liang, D. Li, "A new chaotic algorithm for image encryption", Chaos Solitons & Fractals, 29(2), 2006, pp. 393–399.
- [21] Q. Zhou, K.W. Wong, X. Liao, T. Xiang, Y. Hu, "Parallel image encryption algorithm based on discretized chaotic map", Chaos Solitons & Fractals, 38(4), 2008, pp. 1081–1092.
- [22] Z. Hua, Y. Zhou, C.M. Pun, C.L.P. Chen, "2D Sine logistic modulation map for image encryption", Information Sciences, 297, 2015, pp. 80–94.
- [23] L. Liu, S. Miao, "A new image encryption algorithm based on logistic chaotic map with varying parameter", Springer plus, 2016, doi: 10.1186/s40064-016-1959-1.
- [24] S.E. Borujeni, M. Eshghi, "Chaotic image encryption design using Tompkins-paige algorithm", Hindawi Publishing Corporation Mathematical Problem in Engineering, 2009, 2009, 22 pages.
- [25] H. Khanzadi, M.A. Omam, F. Lotififar, M. Eshghi, "Image encryption based on gyrator transforms using chaotic maps", IEEE 10th International Conference in Signal Processing (ICSP), 2009, pp. 2608–2612.
- [26] Y. Wang, K.W. Wong, X. Liao, G. Chen, "A new chaos-based fast image encryption algorithm", Applied Soft Computing, 11(1), 2011, pp. 514–522.
- [27] H. Khanzadi, M. Eshghi, S.E. Borujeni, "Image Encryption Using Random Bit Sequence Based on Chaotic Maps", Arabian Journal of Science and Engineering, 39(2), 2014, pp. 1039–1047.
- [28] F. Zhou, G. Cao, B. Li, "Design of digital image encryption algorithm based on compound chaotic system", Journal of Harbin Institute of Technology, 14 (Supplement 2), 2007, pp. 30–33.

THE METHODOLOGY OF DATA COLLECTING AND THE REAL TIME SYNCHRONIZATION IN ETL

Zijadin Krasniqi ¹, Enver Ahmeti ², Adriana Gjonaj ³

¹ Ph.D. in Information Systems, Information Technology, Pristina, Kosovo.

² M.sc in Information Technology, Pristina, Kosovo.

³ Professor emeritus, European University of Tirana, Albania.

¹ zijadinkrasniqi@hotmail.com, ² eenver ahmetii@hotmail.com, ³ adriana.gjonaj@uet.edu.al

Abstract-Taking into account the problems and challenges which the Emergency Management Agency (EMA) is facing today, and the way operational data scattered across various institutions is being used, we have set an hypothesis to build a DataMart, which has the capability for deep and detailed data analysis. Here we shall present the step by step modeling of this DataMart, following the Kimball model, which will be useful for detailed analysis by medical, hydro-meteorological and seismic institutes. In collecting, transforming and loading the data we have applied the ETL methods, which are considered as the most critical tasks in building a DWH system. Then the development of a professional modeling in Talend Open Studio for Big Data, which enabled the integration of data from many data sources. In addition to this, we will have the possibility to synchronize and build the data in a way that DWH will be able to communicate in real time, something which has not been possible until now in the field of emergency management

Keywords: *ETL, DataMart, TOS and Big Data, DWH.*

I. INTRODUCTION

Upon building the data mart model, we have to develop all the processes which will enable to gather the information from the original sources, then to transform and load it in its destination [7]. These three processes are considered to be the most complex processes during the phase of building the ETL systems. [5] In accordance with the specific

architecture, the destination of the data will be a data mart built for the process of managing the emergency and natural hazard situations.

In a nutshell, real time data warehouses aim for decreasing the time it takes to make decisions and try to attain zero latency between cause and effect for that decision, closing the gap between intelligent reactive systems and system processes. Our aim is transforming a standard DW using batch loading during update windows (during which analytical access is not allowed) into near zero latency analytical environment providing current data, in order to enable (near) real time dissemination of new information across an organization.[9].

II. DATA EXTRACTING

The first part of an ETL process involves extracting the data from the source system(s). In many cases this represents the most important aspect of ETL, since extracting data correctly sets the stage for the success of subsequent processes. In general, the extraction phase aims to convert the data into a single format appropriate for transformation processing. [4]

To keep the data up to date in data warehouse, data has to be extracted several times in a periodic manner. There are two logical methods for extraction: Full extraction and incremental extraction.

1. Full extraction – In this type of extraction, data from the source systems is completely extracted. As full extraction extracts all the data from source systems there is no need to keep track of changes made in the source system with respect to previous extraction.

2. Progressive extraction – In this type of extraction only the changes made to the source systems will be extracted with respect to previous extraction. This comparison results in large performance impact on data warehouse ETL processes.

There are two physical methods of extraction: Online extraction and Offline extraction.

-Online extraction – The extraction process of ETL connects to source system to extract the source tables or store them in a pre-configured format in intermediary systems, e.g. log tables.

-Offline extraction – The data extracted is staged outside the source systems (ETL Data Extraction Methods – Part Two) [6].

A. Transformation

In the data transformation stage, a series of rules or functions are applied to the extracted data in order to prepare it for loading into the end target. Some data does not require any transformation at all [4].

B. Data loading

The load phase loads the data into the end target that may be a simple delimited flat file or a data warehouse. Depending on the requirements of the organization, this process varies widely. Some data warehouses may overwrite existing information with cumulative information; updating extracted data is frequently done on a daily, weekly, or monthly basis. The timing and scope to replace or append are strategic design choices dependent on the time available and the EMA needs [6].

III. DATA SYNCHRONIZATION

Data synchronization, usually, is needed to keep two or more systems synchronized and up to date. A common reason to synchronize data is system migration (as a result of consolidating or reducing the business etc.). In this situation, it is required to employ two parallel systems for a period of time, sometimes even for several years.

This makes real time data warehouse (RTDW) support a critical issue in this kind of applications. The demand for updated data in data warehouse has always been

very strong. Data warehouse update (the integration of the latest data) traditionally is done using the off-line method. This means that while undergoing updating processes, the users and the OLAP applications cannot use these data. Still, the users are in search of higher updating levels, taking into consideration the fact that the agency of natural hazards operates at 24x7 time shifts. Active data storage refers to a new trend, where the data warehouse is refreshed as many times as possible, because of great user demand to have access to the latest refreshed data [2].

CASE STUDY-MEDICAL INSTITUTE OF KOSOVO (MIK)

A. The tool used for ETL-Talend Open Studio for Big Data

The aim of this chapter is to design and implement ETL processes through the data integration of heterogeneous databases using Talend Open Studio (TOS).

The usage of Talend Open Studio and Big Data (TOS) for designing and semi-autonomous implementing ETL tasks in Java enables a fast way of adopting to new data sources. By doing this we are able to integrate various sources of heterogeneous academic data sample and populate them in one central repository using Talend Open Studio and Big Data (TOS). It is important to ensure the capabilities of open source ETL are equal to any commercial products available, because it will help in implementing DW project with lower costs [1].



Figure 1: The components of Talend Open Studio for Big Data.

Based on the feature of data warehouse system separation, the ETL processes have to be done on a different database, which is independent from the source and the data warehouse. The application of the ETL techniques, by means of the Talend Open Studio

for Big Data component, will be done according to the following steps:

Step 1: Explaining the possibilities of extracting the data from the sources

The source identification is evident, given that the files generated from the platform will serve as the source of the information. If our focus is on the files generated by the Medical Institute of Kosovo (MIK), then these are unstructured files. So, we create a fact table: Tabela_Fact_MIK which contains the data. Five dimension tables connect to Tabela_Fact_MIK and they are the following: Dim_Address, Dim_Blood_group, Dim_Data, Dim_Time and Dim_Health_institu. Dim_Address consist of data regarding dwelling, such as address, residence, contact, phone number, e-mail.

The other table Dim_Blood_group, contains data regarding the blood group, the blood group type and the allergies. Dim_Data contains all types of data, including the following data: the date when the treatment starts, the date the treatment ends. These relate to the medical treatment. While, the other data, such as the Date of Birth, the Day of the Month, the Day of the Year, the Day of the Week, the Number of the Month and the Year show various dates as needed. The table Dim_Time contains time measured by: Second, Minute, Hour, Departure Time, and Arrival Time. The table Dim_Health_Institu contains specific data on: Name and Place of Institution, Department, Unit.

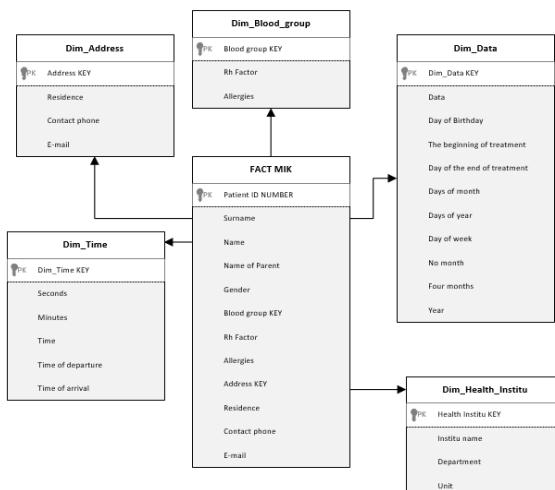


Figure 2: It shows Entity Relationship Diagram of target systems.

In the following chart we can see the sources of all the dimensions, such as Dim_Address, Dim_Blood_Group, Dim_Data, Dim_Time and

Dim_Health_institu, which are Excelfile. We can see also the other source, the table Fact_MIK, built in ORACLE

Table I.: The DBMS source

#	Source DBMS	Table Name
1	Oracle	Fact_KHI
2	Excelfile	Dim_Time
3	Excelfile	Dim_Blood_Group
4	Excelfile	Dim_Data
5	Excelfile	Dim_Address
6	Excelfile	Dim_Health_Institu

This means that each record type in the source file has an assigned numbers of fields different from other types. We need to build a process based on the records of the Medical Institute of Kosovo (MIK). This process will then implement the data structure.

Meanwhile, another source of information, such as the Civil Registration Agency (CRA), generates structured information, where each record of the generated file has the same content. For our data mart, we are very interested on the information generated by MIK. The data extraction process from the source involves:

- The initial process of extracting information;
- The process of incremental extraction.

The initial process takes place when there is no information in the destination database, while the process of incremental extraction provides the extraction of that part of information which has been changed or added since the last version.

Step 2: Information structuring steps

Developing the structuring of information involves the following steps:

- a) In the first step, we will divide the information in separate models, according to the rules of each type of information. The division will be done based on the rules of information differentiation.
- b) The second step is the proper structuring and storing of data, in order to enable their further manipulation. In our case, the data will be stored in Stage tables, according to the chosen architecture. Based on the logic of transformation, we will set the rules to

structure the unstructured information at our disposal. [7].

c) The third step involves using tJoin and tMap operations in Talend Open Studio. It is most convenient to load the input data schemas into the *Repository*. The Repository maintains a list of input and output record structures (schemas) that can be referenced and maintained globally in the system.

B. tJoin component of Talend Open System

Joining two files consists of combining the fields (columns) of one data source with the fields of another data source whenever a key field matches. Joining is accomplished by the tJoin component in Talend Open Studio.

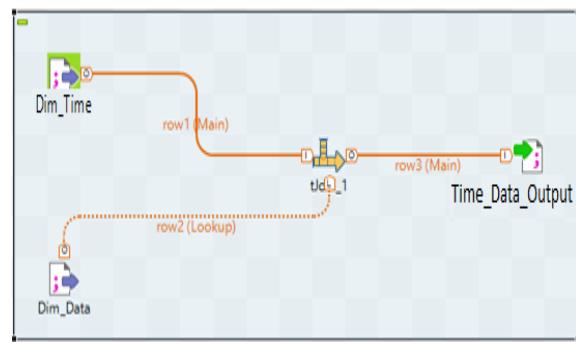


Figure 3: Shows a sample join job designed to combine two data sources

Each data source must be configured to access the appropriate data set according to the requested information in the Basic Settings pane at the bottom. To see the join conditions, double-click on the tJoin component to open the dialog box shown in Figure 4

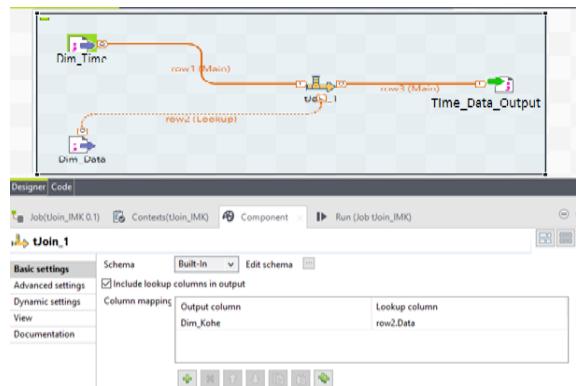


Figure 4: The tJoin component- dialog box

The tJoin dialog box, showing options to specify an inner join and to edit the schema of the join.

If you want to perform a join for areas B and C in Figure 5 (which constitutes a logical right join), just connect the Dim_Data. The common field in each data source (the key) is specified in the definition of the schema in the Repository.

To change the default mapping of input columns, click on the Edit schema box in the tJoin dialog box to display the mapping schema shown in Figure 5 (select View Schema at the prompt, and click OK).

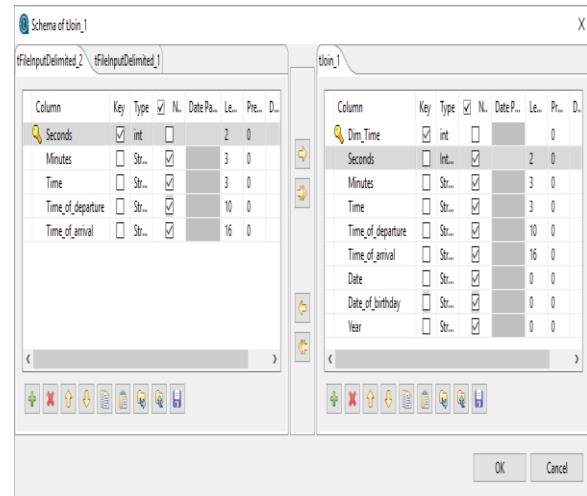


Figure 5: The tJoin field mapping schema for the Dim_Data data source.

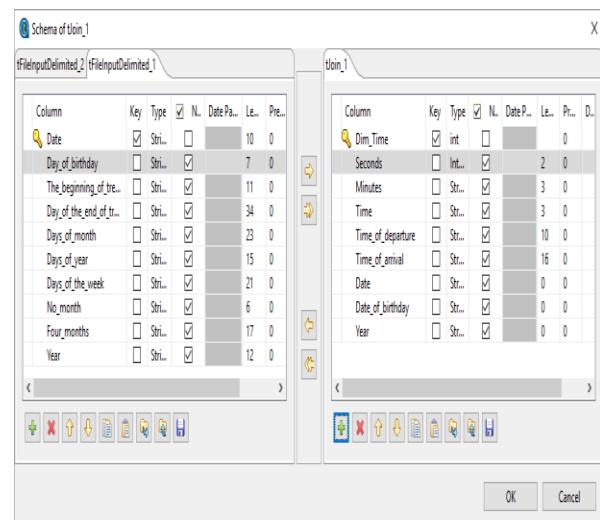


Figure 6: The tJoin field mapping schema for data source Dim_Time.

The goal of the join operation is to add the fields from data source Dim_Time to those of data source

Dim_Data, resulting in a record for a given customer (Time_Data_Output) with all the fields included.

C. tMap component of Talend Open System

Figure 7 shows a job that uses the tMap component to map which input fields relate to each field in the output data structure. Fields can be repositioned in the data records with this component.

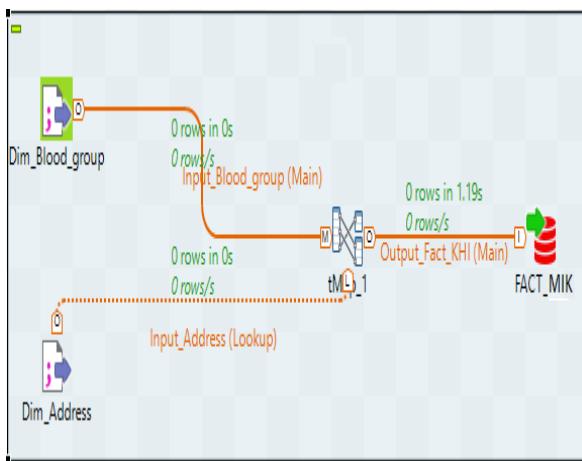


Figure 7. tMap component of Talend Open Studio.

The Figure 7 shows the Talend tMap component for mapping fields to an output data structure. Double-clicking on the tMap_1 component displays the tMap configuration screen in the Figure 8 as below.

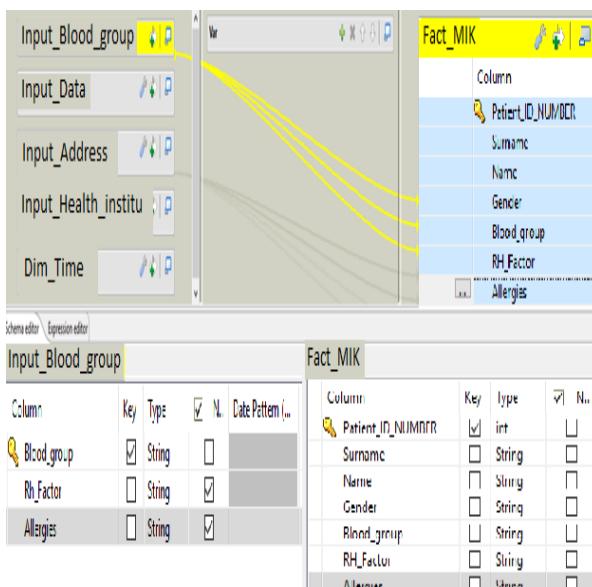


Figure 8. tMap configuration.

Step 3: Specifying the transformation that will happen

After extracting them from the sources, the data undergo the processes of transformation. The transform actions in our case relate to the data extracted from the sources or temporary objects from the Stage zone. The transformations itself can create temporary result that will be stored in the Stage zone. The main transforming processes that will happen are:

- Conversion and normalization that works on both sides of the information to convert the uniform data.
- Joining process, which enables the logical join of equivalent fields from different sources;
- Selection, which reduces the number of the source fields and records

Step 4: Loading the data

During the process of loading the Fact_MIK table, we use the created references. All the references tables needed have to be stored in the memory, so they can be captured again any time a record of this Fact table is processed with the information of the natural key of dimension. Therefore this reference tables are different from the original tables of data mart's Dimension. The Figure 9 below shows the processes that the information undergoes before being loaded in a Fact table [7].

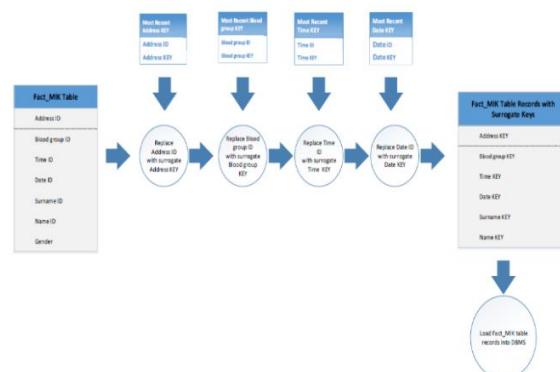


Figure 9: The diagram of loading the table Fact_MIK [3].

The purpose is to show that Talend Open Studio for Big Data (TOS) is capable to perform jobs as any commercial ETL tools.

To prove the proposed framework, we will use dummy data from two different DBMS, which are Oracle and Excelfile to test ETL process in academic environment. Other than that, some record files as Dim_Address, Dim_Blood_group, Dim_Date, Dim_

Time and Dim_Health_institu as text files in Microsoft Excel format have been added as a source data to integrate with DBMS.

Oracle has been chosen as target database because most of enterprises companies are using Oracle in their enterprise applications; in order to load data in the table Fact_MIK we have built the following job in Talend Open Studio for Big Data, as shown in the Figure 10

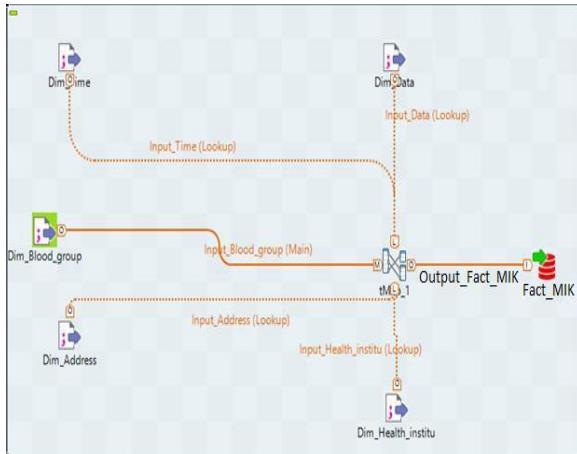


Figure 10. Talend's Job for loading the table Fact_MIK

Meanwhile, in Oracle, a table known as Fact_MIK is created to store the records of Medical Institute of Kosovo, including the patients' number, surname, name, maid's surname, gender, blood group, Rh factor, allergies, location, address, contact number, email, the prescribed medicaments.

Whereas Microsoft Excel files contain data about time, address, date, blood group and medical institution. [1].

During the loading process, we can add many more elements of information, such as the case of the fields after the planning, like surname, name, gender, needed to keep in this table the record's loading time in the database, as shown below.

This added element are not present in the source, but are generated by the system via standard functions of Talend Open Studio for Big Data (TOS).

The abovementioned fields serve to manage the loading process of information or to make examinations in case there might be any problem. [7].

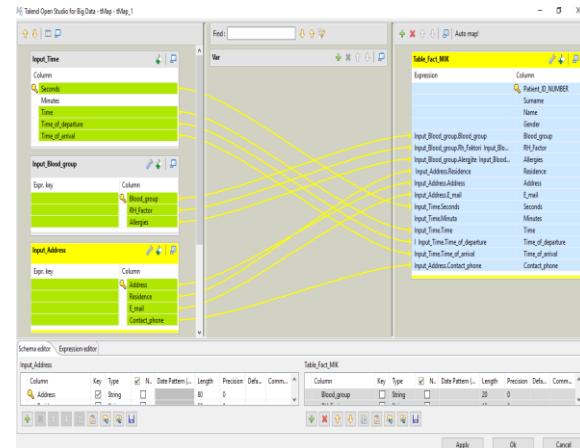


Figure 11 TOS Job to add information.

This means that any record in the original source will have an assigned number of fields different from other types. We need to create a precedence process based on the records of MIK. This process will enable the structuring and adding of the data, as shown in the Figure below.

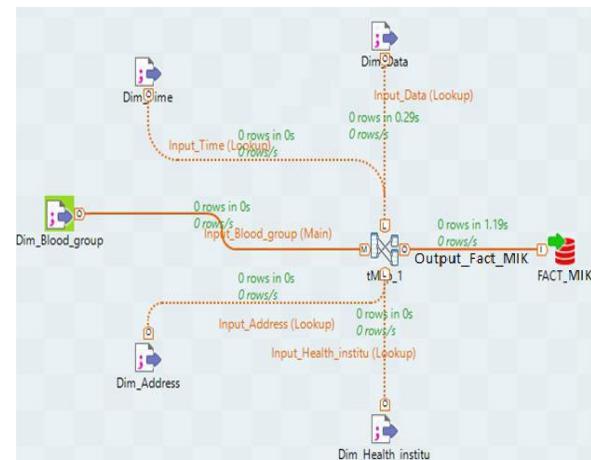


Figure 12. The TOS Job for adding and loading information.

IV. THE RESULTS DERIVED FROM THE ANALYSIS IN THE CASE OF MIK DATA MART

The analysis for the MIK data mart case involves the MIK data mart, which was built to enable exchanging and integrating the data from many dimensions (Dim_time, Dim_blood_group, Dim_address, Dim_data, Dim_health_institu) in a single fact table (FACT_MIK).

Thus, the data mart model above offers high and fast performance in generating the desired results and information, which enables performing analysis within the required deadline. Similar to this data mart model, we can build other data mart models which will

serve to generate information in internet and in small mobile devices.

Taking into consideration the MIK data mart, we can do various analyses regarding the communication form in the Medical Institute of Kosovo. If we take as a cause the requirements of the medical personnel regarding the information relating to the patients, we can easily have the answers for these requirements, simply based on the dimension tables.

Here we can include the time required by the procedures which enable extraction, loading and transforming data. The following table shows the result we got from these procedures:

Table II: Comparing the jobs execution.

#	Job Name	Frequency	Duration
1	Dim_Data	Rows/s=600;ByHour=0;ByMinute=0;BySecond=0.1	00:00:01
2	Dim_Time	Rows/s=189.19;ByHour=0;ByMinute=0;BySecond=0.4	00:00:04
3	Dim_Blood_Group	Rows/s=160;ByHour=0;ByMinute=0;BySecond=0.3	00:00:03
4	Dim_Address	Rows/s=3000;ByHour=0;ByMinute=0;BySecond=0	00:00:00
5	Dim_Health_Institu	Rows/s=6;ByHour=0;ByMinute=0;BySecond=0	00:00:00
6	Fact_MIK	Rows/s=1379.41;ByHour=0;ByMinute=1;BySecond=0.6	00:01:06

Looking at the above results, it is evident that Dim_Data, from 600 rows/s are loaded in 0.1 second, Dim_Time, from 189 rows/s are loaded in 0.4 seconds, Dim_Blood_Group, from 160 rows/s are loaded in 0.3 second, Dim_Health_Institu, from 6 rows/s are loaded in zero second. Thus, all the data from all the dimensions is loaded into Fact_MIK table with 1397 rows/s at a total time of 1.6 min. The MIK data mart case depends on the update of this data warehouse (DWH) within a specific time interval (ore in continuity). Failure to fulfill that responsibility puts the reliability and dependability of the data warehouse MIK in question. Therefore, the data warehouse cannot succeed without efficient and persistent data feeds. [8]

CONCLUSION

The fact is that the Emergency Management Agency possesses a large amount of data, which can easily go up to many megabyte or gigabyte per day. In fact, it is very difficult to run an effective update of this data, so through theoretical explanation based on literature and scientific facts, we have been able to apply real time data synchronization. The ETL process extracts data from the source systems, transforms the data according to the rules of the institution and assigns the

results to a target DWH. Talend Open Studio and Big data software, which enables the integration of data from many different sources, providing a positive example of integration and data modeling.

REFERENCES

- [1] A. Azwa, W. Abdul, H. Nazi rah, “integration of heterogeneous databases in academic environment using open source ETL tools”, February 4, 2015
- [2] A.Hoffer, V. Ramesh, H. Topi “Modern database management-Tenth Edition”, Pearson, New Jersey, 2011.
- [3] C. Johnson, “Surrogate Key in Data Warehouse, What, When and Why”, 2013.
- [4] ETL “Extract transform load “May 12, 2015.
- [5] E. Johnson, J. Jones, “Building ETL Processes for Business Intelligence Solutions Built on Microsoft SQL Server”, March 12, 2015
- [6] K., Kakish Th. Kraft , “ETL Evolution for Real-Time Data Warehousing”, New Orleans Louisiana, Vol. 5,2012, pp 1-12.
- [7] L. Juliana,” Studying data warehouse systems and building a reporting model based on business intelligence technology”, Tirana, 2014
- [8] R. Kimball, J. Caserta, “The Data Warehouse ETL Toolkit-Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data”, Wiley Publishing, USA, 2004.
- [9] R. Santos, J. Bernardino, “Real-Time Data Warehouse Loading Methodology, ACM, Portugal, 2008.

Conceptual Model of Intelligent Collaborative Educational System: Possible Solutions

Sabina Katalnikova

Faculty of Computer Science and Information Technology
Riga Technical University
Riga, Latvia
sabina.katalnikova@rtu.lv

Leonids Novickis

Faculty of Computer Science and Information Technology
Riga Technical University
Riga, Latvia

Abstract - In modern society, an increasingly important role is played by knowledge workers whose professional competence is one of the basic resources of the company and organisation evolution. In order to be competitive, knowledge workers have to keep up with new knowledge, which makes the process of continuous education of employees as one of the priorities in the staff management. One type of lifelong education - collaborative education - provides benefits for cognitive achievements. The Article deals with the characteristics, basic tasks and the operational process of collaborative education. The conceptual model of intelligent collaborative educational system is presented; the functions of system basic models and the scheme for setting up the training scenario are described. Besides, the focus is on the group formation issues.

Keywords - collaborative education; competence approach; intelligent collaborative educational system; knowledge worker

I. INTRODUCTION

In today's society, no one can call into question the effect of new technologies on the educational system. In this new reality, at any level of education, the study process should be organised in such a way that graduates become the so-called knowledge workers in the full sense of this term [7]. This means that the training becomes more learner-centred, providing training opportunities for individual and professional development purposes, and educational institutions seek for flexible and dynamic response to changes on the labour market and in daily life.

One of the characteristics of the 21st century is the transition to the knowledge economics and the transition of information to the source of economic growth and profit. The staff and their knowledge as well as their growing professional competence have become the basic resources of organization development. In order to be competitive, it is necessary for knowledge workers to continuously acquire innovative knowledge. This makes the process of staff continuous training as one of the priorities in the functions of staff management in modern organisations. Thus lifelong education is becoming increasingly important.

According to [14], lifelong education is the process of personal, social and professional development of the individual throughout their lifetime to improve the quality of life. It is a

comprehensive and uniting idea which includes formal, beyond-formal and non-formal education.

Lifelong education is a continuous process in which a person is involved from youth to old age, it requires flexibility of mind, the desire to complement and broaden the knowledge as well as the desire for continuous development as a full and versatile person, and it enables an individual to adapt to the changes on the modern labour market. The rationale for adults to engage in lifelong education can be varied, both internal and external. Internal motives refer to the wishes of the person to acquire new competences; the external motives are the various demands of person's workplace.

On the other hand, information and communication technology opportunities and capacity are constantly increasing, leading to a continuous increase in the volume and variety of service activity in any area, including education. Therefore, the concept of collaborative educational system has appeared in the modern society. On the modern labour market, there is an increasing demand for creative, responsible, dynamic and well-prepared employees who can quickly adapt to changes, who know how to cooperate with other employees, who are open to different cultures, with flexible thinking, etc. [1]. In this context, collaborative education is becoming increasingly important.

II. COLLABORATIVE EDUCATION WITHIN THE FRAMEWORK OF COMPETENCE APPROACH

Collaborative education can be defined as a training environment in which students participate with a purpose to tackle the challenges together [16]. Collaboration activities enable students to provide explanations based on their understanding, which assist them in developing and reorganising their knowledge [3]. Collaborative education provides the following advantages [13]:

- the development of critical thinking skills,
- joint creation of knowledge and meaning,
- deliberation,
- transformational learning.

There are four basic elements of collaborative education [5]:

- positive interdependence – students must be fully involved in the work and strive to achieve the common objective of the group; in this case, each team member has an assignment, and the student understands that the execution of an individual task affects the operational effectiveness of the whole group;
- “eye to eye” interaction – each member contributes to the success of other participants. Students explain to each other what they have learned and help each other to understand and complete the task;
- social skills – collaboration in the field of education to be successful, everybody needs to develop effective communication types;
- group evaluation – in the process of collaborative education, the group often has to assess operational efficiency and to decide how to increase it; in order to achieve a significant improvement, students have to possess two features: each of them is working for the group aim, and success depends on everybody's results; the contribution and responsibility of each student are recognized in the assessment.

Another topical issue at present: competence approach in education. For a long time, the qualifications of employees were defined by their knowledge, based on what the person was studying and how long. However, at present a knowledge-based approach is not topical because nowadays, the specialist's abilities after studying matter as well as their ability to respond adequately to the changes in the situation and to act creatively.

Another cause of the decline in the importance of knowledge is linked to global Web development because at the moment, information is available to anyone without any expert mediation.

The idea of knowledge, skills and abilities makes it possible to act by analogy with the model, but the competence is based on independent activities at the basis of universal knowledge. As a conceptual definition, it is possible to adopt the term “competence” as defined in the European Union project TUNING: The term “competence” includes knowledge and understanding (theoretical knowledge in an academic field, ability to know and understand), the knowledge of how to take action (practical application of knowledge in particular situations) and the knowledge of how to exist (values as an integral part of life in a social context).

This European study in the field of the assessment of qualifications and competence is a part of international project Tuning Educational Structures. The project is based on the introduction of the suggestions of Bologna Process, the harmonisation of the parameters of education programmes at all levels as well as an improvement in the interaction of higher education and entrepreneurship being at the basis of competence approach [18].

The basic principles of competence approach in education are as follows:

- constant updating of the study programme in accordance with the requirements of the labour market;
- development a study programme according to the individual student's needs;
- necessity of serious motivation for the acquisition of theoretical knowledge and practical skills in the chosen field;
- evaluation of the will to study (strengths and weaknesses of the student, individual characteristics, etc.).

How should competences be created? In general, the development of one competence must correspond to a number of different disciplines, grouped in a specific way (of course, there are situations where one competence is generated within the course of one study course, or even within a part of the course). A set of study courses constituting a single competence or a group related competence is referred to as a module. The formation of one competence can include studying different disciplines, participation in student conferences, various kinds of traineeship and unassisted work.

As it has been said, collaborative education is a topical issue in today's society, the main economic resource of which is knowledge and where knowledge workers are of great importance. The term “knowledge workers” was introduced by Peter Ferdinand Drucker – a scientist, economist, educator (USA) - in 1959 [12] to describe an ever-increasing number of employees working mainly with information or using knowledge. The key aspects of the operations of knowledge workers are innovation, responsibilities and collaboration that have led to the great importance of collaborative educational systems for knowledge workers. The above approach to the implementation of the concept of lifelong learning is therefore complemented with a need to take into account the characteristics of knowledge workers.

In general, the following approach is proposed to implement the concept of lifelong education (see figure 1):

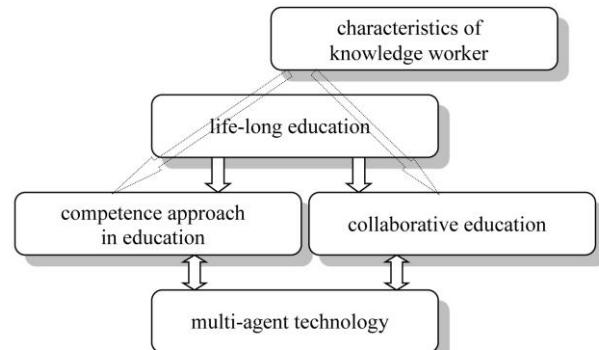


Figure 1. Proposed approach to the implementation of an idea for the lifelong education of knowledge workers

III. SYSTEMS OF COLLABORATIVE EDUCATION

A distinctive feature of collaborative educational systems is the development of domain knowledge, the use of individualised education and training strategies based on the model of the user. The benefit of collaborative educational systems is an advantage of independence from a student placement, but the difference between the classes of users is an important problem, as the system should be used for a very diverse set of students.

Systems of collaborative education possess the following characteristics:

- the aim of the system is to provide students with adequate support in the process of problem solving as an instructor-human would have done;
- any system of collaborative education is based on a user (student) model, taking into account the set of characteristics of the adults as well as the current level of professional competence and knowledge;
- at the basis of studies, there is the underlying study programme based on the student characteristics and needs, the purpose of education and the required range of competencies;
- education is geared to the needs of employers;
- during studies, there are developed skills for the user to apply knowledge for specific practical tasks;
- self-motivation is of great importance in such systems;
- education takes place in collaboration, promoting equal partnership.

In general, the key objectives of collaborative educational system are considered to be:

- the development of a model of competence based on existing standards and labour market requirements;
- student registration and development of his or her personal space;
- the development of the current model of each participant based on a deep and thorough diagnosis and a comparison with the competence model with a view to determining the goal of the education process of each participant;
- development of a training strategy (programme) at the basis of the goal of student education process;
- formation of collaborative education groups, taking place separately for each study course module based on the current status and personal characteristics of students;
- developing the content (scenario) based on the existing training strategy and the groups formed;
- formation of teaching staff and auxiliary staff team, based on employee characteristics and study programmes;

- determination of labour-intensity and labour input of the training scenario;
- implementation of training, support for maintaining the training process, formation of teaching-methodology group;
- implementation of control, updating of the student model, determination and evaluation of study results.

The operational process of the collaborative educational system is summarized below (figure 2).

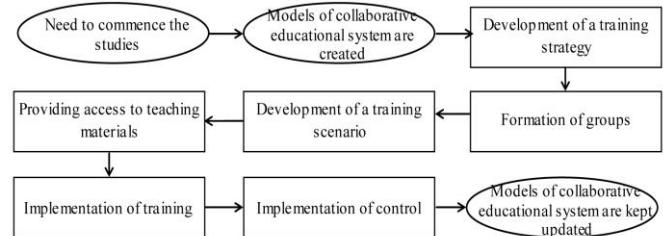


Figure 2. Process of the collaborative educational system

As part of this process, an inspection is carried out as to whether each unit established corresponds to the model of the student (specialty, training objectives and the level of knowledge, skills and competences) and the model of competence. If there is conformity, the unit is considered to be created, otherwise the process will recur.

IV. CONCEPTUAL MODEL FOR THE INTELLIGENT COLLABORATIVE EDUCATIONAL SYSTEM

The term “intelligent system” is used in the field of artificial intelligence and identifies a system using artificial intelligence with a view to providing better support for system users. There is no uniform definition of intelligent system at present, for example [20] gives the following explanation: “The technical or software system, the operational resolution of which is traditionally regarded as creative and belonging to a particular item of understanding, the knowledge of which is kept in the system memory; an intelligent system consists of three main blocks: the knowledge base, solver (output mechanism) and an intelligent interface”.

Mostly intelligent systems are knowledge based systems and focus on processing knowledge (not on processing data or information). Therefore, the intelligent system is a system based on knowledge, i.e. a domain model, described in the language of knowledge interpretation (ultra-high level programming language, close to natural language) [21].

Intelligent systems are widely used in different fields to meet versatile challenges. One of such tasks is planning – organisation of a system with a view to achieving a defined objective taking into account certain restrictions. Intelligent educational systems also belong to these systems.

A possible conceptual model for the intelligent collaborative educational system, namely, the main subsystems and data flow between them, is presented in figure 3:

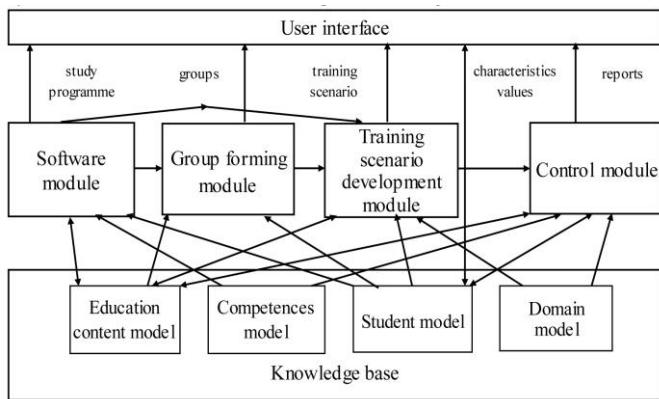


Figure 3. Conceptual model of the intelligent collaborative educational system

The education content model includes the strategies (regulations on the behaviour of the system) of study programme and study content (scenario) development and management. The student model contains information about the student, the set of characteristics (personal data, level of knowledge, etc.). Competence model: a set of required competences based on industrial requirements comparing current and acquired competences of students. Domain model includes study objects (study courses, tasks, tests, control questions, etc.).

The models are in accordance with the base modules, the rest of modules provide ancillary functions. This could be: module design and analysis module, module for the implementation of training, module for knowledge base validation (correctness check) module, etc.

We will review the functions of the base modules.

User interface:

- registration of users and development of personal space;
- access to the knowledge base;
- interaction with other modules;
- input of current values for the assessment of student competences, values for student characteristics, etc.

Module for programme development:

- comparison of the value of student's competence rating values with the value required for the assessment of competence (i.e., competence model);
- determination of the study programme (training strategy), taking into consideration the student's characteristics as well as the difference between the value of student competence assessments and the values required according to the assessment of competence, i.e. the definition of a set of study modules for each student.

Group formation module:

- based on the established programme for each student, his/her personal characteristics and current situation, collaborative groups are created for module studies of each study course;
- checking the adjustments of groups created for each study course module. If a group is not formed, the students outside groups are united in groups, not taking into account the characteristics and present situation.

Training scenario development module:

- sequence planning for study course modules;
- implementation of training, support for the maintenance of the study process.

Control module:

- implementation of the control of acquired knowledge;
- upgrading of the student module;
- generation of management solutions appropriate for study process to achieve study objectives;
- determination and evaluation of study results.

All the activities of the system are carried out, taking into account the limitations of the resources distributed, and directly linked to the student. Development of model training scenario could therefore be presented as shown in figure 4.

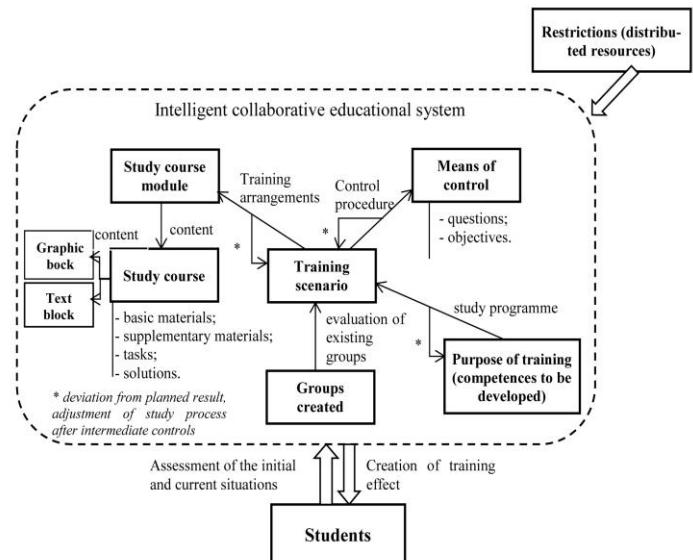


Figure 4. Diagram for development a training scenario

The following concepts are highlighted in figure 4:

- the text and graphical block is the minimum unit of presenting information in the text and in the graphic form respectively (e.g. paragraph, sentence of the text block, etc.);
- the study course allows the combination of different information in the form of text and graphical block in various structures;

- a study module is a set of study courses (it may also be traineeship) which has a logical completeness and that is intended to a certain competence, i.e. study of interrelated courses.

V. QUESTIONS RELATED TO USER GROUP FORMATION OF INTELLIGENT COLLABORATIVE EDUCATIONAL SYSTEM

Collaborative education provides benefits for cognitive achievements, but their effectiveness depends on such factors as the previous knowledge of group participants, team composition and quality of explanations [8]. Without adequate prior knowledge, students fail to provide high-quality explanations or properly understand the perspectives provided by other members of the group [10].

One of the problems related to collaborative educational system: how to divide students into study groups and according to which criteria. One approach: dividing into groups based on the level of knowledge and current competencies of each member, taking into account also his/her cognitive and social capacity.

There are also a number of approaches in determining the optimum number of participants in the group. It is most often recommended to divide the participants into as small groups as possible, 2-5 people in a group [6, 9].

In the research [4], the author offers a review of literature sources related to research of collaborative education. It has been noted that, within the framework of collaboration, both interaction models and the success of collaboration vary across groups with different levels of participants' capacity. In the research, studies were considered in which groups with both a high spectrum and a narrow spectrum participated (where there were students with high – mean capabilities and mean – low capabilities). In the groups with a wide spectrum of abilities, high and low capacity participants represented the teacher-pupil relationship, while the medium level capacity participants stayed away. In the groups with narrow capabilities, medium capacity participants worked much better. In these groups, all the students had a tendency to participate actively in the work, and medium and low capacity participants showed a higher achievement. The research has shown that collaboration has a strong impact on student learning outcomes, especially as it relates to participants with low skills, and the groups with a narrow difference in capacity levels are the most successful ones.

Development of an efficient collaborative educational system includes sub-tasks: how to choose and quantify appropriate characteristics to develop a system user model to divide users into optimal groups for study purposes [2].

The research [11] provides a methodology for determining the level of competence of knowledge workers: the level of competence of staff members is calculated by means of an audit procedure based on different methods and techniques of competence analysis. From an audit point of view, each competence includes the title and attributes set to define it. Each attribute of a particular worker is assessed in some way (for example, through a survey or an interview). The collected

attributes make it possible to calculate the staff member's level of competence.

Each employee i possesses a set of competences characterized by competence vector $Com_i [com_{1,i}, com_{2,i}, \dots, com_{n,i}]$ where n is the number of competences and the value $com_{n,i} > = 0$ is the initial assessment of competences carried out with an audit procedure. In addition, value $com_{n,i}$ can be changed after the loss or transfer of knowledge, training or other knowledge related processes.

Another question which has to be addressed arises: is competence a binary property or not? In the field of education, competence may be assessed by scale, with degree or "size". In the research [15], it is noted that competence may consist of several subdomains and may include the use of a number of related elements, so that a lower level competence can be considered binary, whereas a higher level competence - able to be partially completed. For example, the level of competence may be normalized and linked to the audit procedure of a specific employee: beginner (0-0,2), initiator (0,2-0,4), trainee (0,4-0,6), skilled employee (0,6-0,8), expert (0,8-1) and master (1) [17].

VI. CONCLUSION

There is no doubt that major changes in higher education are currently taking place. Educational institutions offer ever more lifelong education opportunities for lifelong education and adult education. The main aim of this activity is to ensure the availability of quality and efficient education for all. In the circumstances where demand for such opportunities is constantly rising, the importance of the intelligent collaborative educational system and development issues associated with it should not be exaggerated, since the development of education should take place directly from the "closed" education model to an open one based on the use of a virtual education environment as the main means of communication between students and educators.

The article deals with the concept and core functions of the concept of the intelligent collaborative educational system, as well as the conceptual model of the intelligent collaborative educational system has been promoted, which might be helpful to the developers of this type of intelligent systems.

REFERENCES

- [1] A.C. Damian, M.Georgescu. Collaborative and Self-directed Learning in a Virtual Campus Environment: a Potential Solution for Our Years? EcoForum, Vol. 3, No. 1 (2014). [Online]. Available at: <http://ecoforumjournal.ro/index.php/eco/article/download/52/52> [Accessed: October 25, 2017].
- [2] C. Long, Y. Qing-hong. A group Division Method Based on Collaborative Learning Elements // Proceedings of the 26th Chinese Control and Decision Conference, 2014, pp. 1701-1705.
- [3] C. Van Boxtel, J. Van der Linden, G. Kanselaar,. Collaborative Learning Tasks and the Elaboration of Conceptual Knowledge. Learning and Instruction, 10(4), 2000, pp. 311–330.
- [4] E.R. Lai. Collaboration: A Literature Review. Research Report, 2011, 49 p. [Online]. Available: images.pearsonassessments.com/images/tmrs/collaboration-review.pdf [Accessed: August 10, 2017].

- [5] H. Brown, D.C. Ciuffetelli. Foundational methods: Understanding teaching and learning. Toronto: Pearson Education, 2009. – 584 p. ISBN: 978-0558371968.
- [6] J. Bernacki, A. Kozierkiewicz-Hetmańska. The Comparison of Creating Homogeneous and Heterogeneous Collaborative Learning Groups in Intelligent Tutoring Systems// Proceedings of 7th Asian Conference of Intelligent Information and Database Systems, 2015. Part I, pp. 46-56.
- [7] J. Grundspenkis. The Conceptual Framework for Integration of Multiagent Based Intelligent Tutoring and Personal Knowledge Management Systems in Educational Settings // Workshops on Business Informatics Research: International Workshops and Doctoral Consortium. – 2011. - pp. 143-157.
- [8] J. Janssen, F. Kirschner, G. Erkens, P.A. Kirschner, and F. Paas. Making the black box of collaborative learning transparent: Combining process-oriented and cognitive load approaches. *Educational Psychology Review*, 22, 2010, pp. 139–154.
- [9] J.K. Olsen, D.M. Belenky, V. Aleven, M. Ringenberg, N. Rummel, J. Sewall. Authoring Collaborative Intelligent Tutoring Systems // Proceedings of the Workshops at the 16th International Conference on Artificial Intelligence in Education, 2013, p.10.
- [10] N.M. Webb, K.M. Nemer, A.V. Chizhik, and B. Sugrue. Equity Issues in Collaborative Group Assessment: Group composition and Performance. *American Educational Research Journal*, 35(4), 1998, pp. 607–651.
- [11] P. Różewski, J. Jankowska, P. Bródka, R. Michalski. Knowledge Workers' Collaborative Learning Behavior Modeling in an Organizational Social Network. *Computers in Human Behavior*, Volume 51, Part B, October 2015, p. 1248–1260.
- [12] P.F. Drucker. Landmarks of tomorrow. Harper Business, USA, 1st edition. -1959. - p. 270.
- [13] R.M. Palloff, K. Pratt. Collaborating Online: Learning Together in Community. San Francisco, CA: Jossey-Bass. 2004, p. 128. ISBN: 978-0-7879-7614-9.
- [14] R.N. Dave. Foundation of Lifelong Education: Some Methodological Aspects. Hamburg, Pergamon. 1976. – 388 p. ISBN: 978-0-08-021192-3.
- [15] S. Grant, R. Young. Concepts and Standardization in Areas Relating to Competence // *International Journal of IT Standards and Standardization Research*, 8(2), 2010, pp. 29-44, DOI: 10.4018/jitsr.2010070103.
- [16] S.D. Teasley, J. Roschelle. Constructing a Joint Problem Space: The computer as a tool for sharing knowledge. In Lajoie, SP and Derry, SJ (eds), *The Computer as a Cognitive Tool*. Hillsdale, NJ, USA. Erlbaum, 1993, pp. 229–258.
- [17] T. Farrington-Darby, J.R. Wilson. The nature of expertise: A review // *Applied Ergonomics*, 37(1), 2006, pp. 17-32.
- [18] Tuning Educational Structures in Europe. [Online] Available at: <http://www.unideusto.org/tuningeu/> [Accessed: August 2, 2017].
- [19] J. Griesbaum. Mehrwerte des kollaborativen Wissensmanagements in der Hochschullehre: Integration asynchroner netzwerkbasierter Szenarien des CSCL in der Ausbildung. Hülsbusch, W, 2009. 480 p. [In German].
- [20] A.N. Averkin, M.G. Gaaze-Rapoport, D.A. Pospelov. Explanatory Dictionary on Artificial Intelligence. M.: Radio and communication, 1992. – p. 256 [In Russian].
- [21] G.V. Ribina. Basics of Development of Intelligent Systems. M.: Finance and Statistics, 2010, p. 432 [In Russian].

Develop a distributed Intrusion detection system based on Cloud Computing

Khoi Nguyen

Datic Lab, DaNang University
DaNang, Viet Nam
ntkhoi@ dut.udn.vn

Trang Dang Thi

IT Faculty, PhamVanDong University
QuangNgai, VietNam
thuytrang1401@gmail.com

Abstract— An intrusion detection system (IDS) is a hardware device or software application that monitors network and/or system or host activities for malicious activities policy violations. IDS has been used widely in network systems to detect malicious behaviors which can harm system or computers. Snort is an open source network intrusion detection and prevention system. It can analyze real-time traffic analysis and data flow in a network. It is able to detect many different type of attack. In this paper, we introduce a Cloud IDS model as a solution to implementing a network Snort based on cloud computing. The Cloud IDS provides network IDS as a service over the internet which can be simple in deployment, maintenance, scalability without investing in the new infrastructure.

Keywords: IDS; Snort; OpenStack; Cloud Computing

I. INTRODUCTION (HEADING 1)

In recent years, Cloud computing technology has developed and become popular with many providers like Amazon, Microsoft, Google, etc. The abilities to share resources among several users, dynamic scalability, and pay-per-use basis etc. bring the convenience to users. The network security is one of the most important parts of a network system, but the traditional technologies like firewall, encryption... is not good enough to protect the network against many new attacking mechanisms. The IDS has been proposed and widely used as an efficient security system to identify internal and external malicious behaviors. Although considered as an efficient tool to protect a network system, IDS has to analyze all packets in network to detect malicious behaviors. Therefore, the processing time will be high and may lead to congestions, and packets may be delayed or dropped. Besides, because of the complexity of deployment and operation, companies must invest in infrastructure and personnel training.

Intrusion detection system plays an important role in the security and perseverance of active defense system against intruder hostile attacks for any business and IT organization. IDS implementation in cloud computing requires an efficient, scalable and virtualization based approach. In cloud computing, user data and application is hosted on cloud service providers remote servers and cloud user has a limited control over its data and resources. In such case, the administration of IDS in cloud becomes the responsibility of cloud provider. Although the administrator of cloud IDS should be the user and not the provider of cloud services. In the paper [1][6] authors have proposed an integration solution for central IDS management

that can combine and integrate various renowned IDS sensors output reports on a single interface. The intrusion detection message exchange format (IDMEF) standard has been used for communication between different IDS agent.

In this paper, we introduce Cloud IDS which is built as a SaaS (Software as a Service) to provide IDS as a service to cloud based user through the internet. This model helps users reduce costs and becomes simple in deployment and operation.

II. PROPOSED AN CLOUD IDS MODEL

A. Intrusion Detection System (IDS)

Intrusion Detection System (IDS) is a system that analyzes packets in a network to detect malicious activities or policy violations and produces reports to the administrator. The IDS can identify and detect malicious behaviors from inside (agents, customers...) or outside (hacker...)

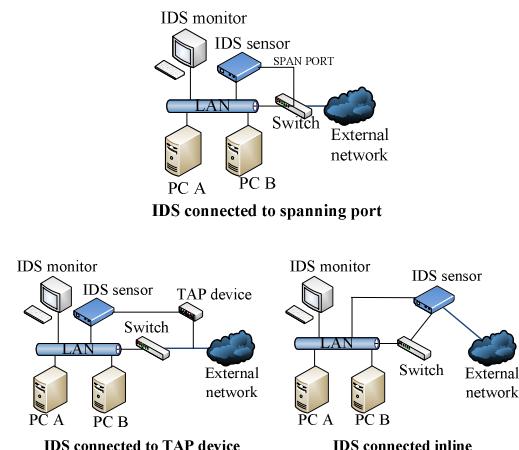


Figure 1. Network IDS deployment strategies

Beside many advantages, there are some challenges that face traditional IDS. For example, the user must invest in the new infrastructure, personnel training, etc. in order to deploy a network IDS. The IDS must capture and analyze all packets in the network to identify malicious activities. Therefore, in a large network it can lead to congestions and packets may be delayed or dropped.

B. Cloud computing

Cloud computing is an innovative computing model in which resources are provided as a service over the internet based on the user demand. Users of cloud computing services can request, scale up, and use services without intervention of the provider. User doesn't need to invest in infrastructure, reduce maintenance cost and just pay only for the resources that they use. Cloud computing providers offer their services according to several fundamental models: IaaS, PaaS, SaaS [1][5].

C. Proposed Model

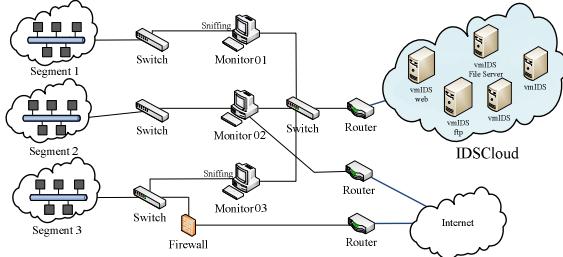


Figure 2. Proposed IDSCloud Model

In this section, we propose a model with the goal of providing IDS as a service to users in order to solve the problems of the traditional IDS and inherit advantage features of cloud computing. We named it CloudIDS.

In this proposed model, the user integrates a light weight sensor into the protected network and installs a central management application to monitor reports. The Detection Unit is rented from IDSCloud service. Based on specific demand, users can rent as many Detection Units as they need, customize the signature database. The User is the only one that completely possesses the analysis results. This model promises to provide users with an IDS with easy deployment, no costly investment in infrastructure, less maintenance cost, scalability...

Cloud IDS consists of four components:

- Capture component: the software or hardware is integrated into the user network; the purpose of this component is to capture packets in network. According to where this component is placed, it can capture packets in the whole network or network segment.
- Transportation component: this component consists of two elements, the transmitter and the receiver. The transmitter is placed at users' side and send collected packets to receiver, the elements which are placed at IDSCloud service. A Secure connection is established between two elements in order to protect the transmission of data from the external intrusion

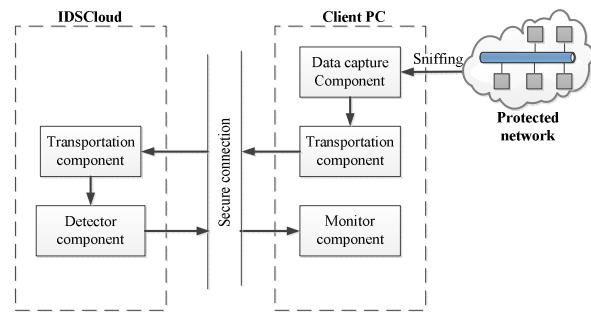


Figure 3. IDSCloud Architecture

- Detector component: The most important part of IDSCloud, a detector component analyses packets which the Transportation component received in order to detect malicious activities. Usually, Analysis process consists of four stages: packets decoder, preprocessor, engine detection, logging and alert system.
- Monitor component: This component is a real-time network monitoring tool. In order to guarantee that users possess analysis results and data does not lose when the detector unit IDSCloud is terminated by user, the monitor component is placed at user's side and receive report and alert from detector component via a secure connection path.

IDS Cloud Requirements are as follows:

- Capture component: In order to protect user's network completely, IDS Cloud needs to monitor all the traffic on protected network; therefore, the capture component must be placed where it can capture all packets in protected network such as Test Access Port (TAP), Switched Port Analyzer(SPAN). It has the ability to work with high-speed networks and user can apply filter to it to reduce analyzed packets.
- Detector component: because the purposes of each user is different, this component should have ability to customize the detector to optimize the user's network data analysis. Besides, like other IDSs signatures database can be updated to improve detection ability.
- Secure connection: IDSCloud's user needs a secure connection to send and receive data with IDSCloud service. The secure connection can protect user's data from external intrusion.

In the next part, we are going to implement and evaluate our proposed IDSCloud model in local test.

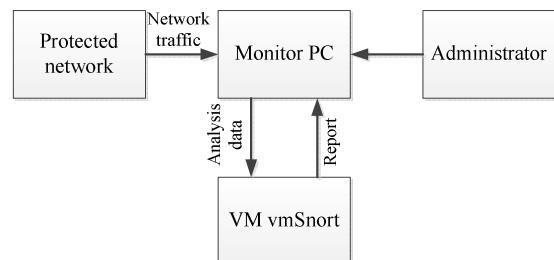


Figure 4. IDSCloud activity diagram

- 1) Infrastructure as a service(IaaS): What IDSCloud provides is a computing resource that has the ability to analyse network packets and identify malicious activities. The computing resource is a virtual machine(VM) with the preinstalled IDS software, it will analyse data received from user and send reports to the central management. Openstack is used to deploy as an IaaS solution and aimed at users such as cloud service providers, goverment, company etc that want to implement large-scale public cloud computing, private cloud computing or hydrid cloud computing.
- 2) IDS software: We chose Snort as the IDS software in the proposed model. Snort is one of the most popular free and open source IDS developed by Sourcefire. Snort is a signature-based NIDS but it can be expanded to use anomaly based detection plugins as well. The signature database is updated frequently and Snort can analyse protocols, header packets to identify many attack methods like buffer overflow, stealth ports scan, CGI attacks... In this experimental scenario, Snort is configured to use DAQ_pcapspooler as Data Acquisition library in order to monitor received folder and start analysing data file as soon as it arrives.
- 3) Secure Connection: The data transmission between clients and IDSCloud service is sent through Virtual Private Network(VPN) in order to ensure data is kept safe. Besides, users can use VPN to manage, config virtual machines at IDSCloud service. User's computer is connected to IDS VM through a VM named vmGateway.

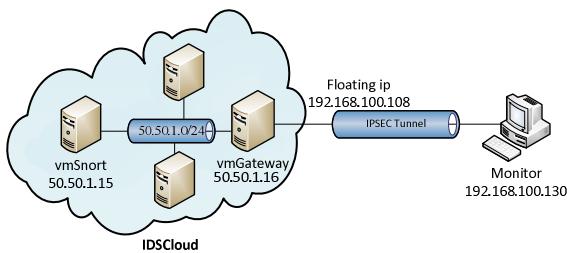


Figure 5. VPN Configuration with StrongSwan

III. EXPERIMENTAL EVALUATION

In order to evaluate the performance of the IDS Cloud model we have built at section 5, we propose a attack detection scenario and compare the performance result between IDSCloud and the traditional IDS. In figure 6 illustrate the experimental scenario.

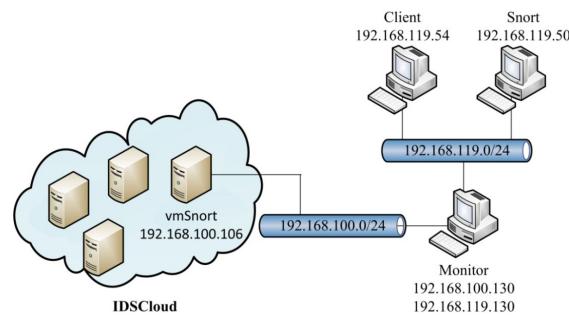


Figure 6. Experimental scenario

The evaluation environment consists of 4 PCs, the PC Monitor using IDSCloud service and PC Snort using IDS snort to protect network segment 192.168.119.0/24. The PC Client (192.168.119.54) will attack the protected network. The VM vmSnort is a virtual machine on IDSCloud. The Monitor PC,Snort PC and Client PC have same specification(1 CPU, 768MB RAM, OS Ubuntu 12.10 64 bit). The Snort PC and VM vmSnort use the same Snort version 2.9.6.0 with rule database version 2955 downloaded from Snort website.

The testing pcap data file contains 1.240.251 network packets which trigger snort and generate 3139 alerts with preinstalled rule database. The Client PC send packets to protect network by using Bittwist tool. We run evaluation case twice and evaluate the number of packets captured and alerts generated by two IDS. The results have been shown in Figure 7.



a) Number of captured packets comparison

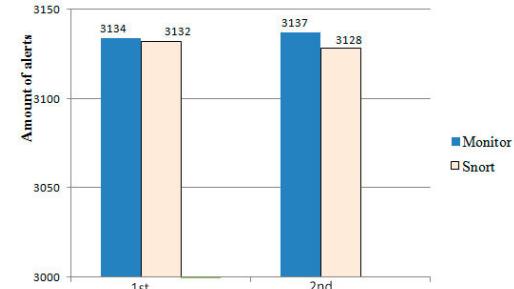


Figure 7. Compare the results between the traditional Snort system and the IDSCloud

The results in Fig 7 show that the IDSCloud can capture more packets than the traditional Snort in the same condition (it also means the less dropped packets) and the result is that the IDSCloud detects more malicious activities.

IV. CONCLUSION

In this paper, we propose an Cloud IDS model as a solution to providing IDS as a service to users. This model has advantages over traditional IDS; it helps users reduce maintenance cost, dynamic scalability and pay-as-you-use feature. In the implementation and evaluation, we deploy IDSCloud model in virtualization environment and evaluate it with pre-captured network traffic file. The results show the effectiveness of proposed model but it is just the first step. There are many factors that can add an extra cost, so for the future works, we will research in the direction of simplifying deployment of Capture Component, optimizing the packet filter to reduce amount of data transferred between client and Cloud IDS, and applying more powerful detection mechanism to Snort as well as evaluating model on real internet environment.

REFERENCES

- [1] Gartner (2013), Forecast Overview: Public Cloud Services, Worldwide, 2011-2016, 4Q12
- [2] Hassen Sallay, Khalid A.Alshafan, Ouissem Ben Fred (2009), A scalable distributed IDS Architecture for High speed Networks, International Journal of Computer Science and Network Security, VOL.9, No.8.
- [3] Kevin Jackson (2012), OpenStack Cloud Computing Cookbook, Packt Publishing.
- [4] Ms. Parag K. Shelke, Ms. Sneha Sontakke, Dr. A. D. Gawande (2012), Intrusion Detection System for Cloud Computing, International Journal of Scientific & Technology Research Volume 1.
- [5] Richard Hill, Laurie Hirsch, Peter Lake, Siavash Moshiri (2013), Guide to Cloud Computing, Springer.
- [6] S. Roschke, F. Cheng, and C. Meinel, "Intrusion detection in the cloud," in 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing. IEEE, 2009, pp. 729–734.

AUTHORS PROFILE



Khoi Nguyen received the B.S. degree in Information Technology from the Danang University of Technology, Vietnam, in 1997, the M.Sc. degree in Computer Science in 2003, and the Ph.D. degree in Computer Science from the Aix-Marseille University in 2010. He now is lecturer at University of Science and Technology, Danang, Vietnam. His current research interests include Distributed System, Network Security.

Trang Dang Thi received the the M.Sc. degree in Computer Science in 2014. He now is lecturer at PhamVanDong University, QuangNgai province, Vietnam. His current research interests include Computer network, Network Security.

A comprehensive approach on RLE and ECC (Elliptical Curve Cryptography) using Mean Square Error (MSE) feature

Ayushi Mathur

Department of Computer Engineering
Government Women Engineering College
Ajmer, India
ayushi1ajmer@gmail.com

Dr. Varun Prakash Saxena
Department of Computer Engineering
Government Women Engineering College
Ajmer, India
vps@gweca.ac.in

Abstract— Elliptic Curve Cryptography (ECC) has various issues related to its security, space and time complexity. To improve the performance of information system, we must need a hybrid system to provide the fast transaction with confidentiality and efficiency. This paper proposes a combinational and comprehensive approach for information transaction. In this direction, we work on a compression algorithm RLE (Run Length Encoding) and ECC to overcome the problem of data storage and security. However, the result of the compression ratio shows that the data compression ratio and the time complexity of encryption process gets reduced. Hence, much better results are achieved under the new implementation. In this proposed approach, a new feature called MSE (Mean Square Error) is added which checks the status of recovered message. Additionally, much better permutation method also used to generate a different private key every time to make ECC more safe and efficient.

The complete improvement and implementation is done on MATLAB R2013a version

Keywords- MSE, compression ratio, RLE, Time Complexity, ECC

I. INTRODUCTION

In the existing scenario data complexity or the data storage problem got reduced by using RLE algorithm and data security got increased by using permutation method for private key generation [2]. This paper calculates the compression ratio of the compressed data with the original data and also calculates the time taken for encryption process. Encrypted data is also decrypted to recover the original data. Mean Square Error (MSE) is calculated to recheck whether original data got recovered or not. Compression ratio is mentioned in percentage and denotes the extent to which the original data got compressed after using RLE algorithm. Below two formulae shows how Compression ratio and MSE is calculated.

$$\text{Compression ratio} = \frac{\text{Length of compressed data}}{\text{Length of original data}}$$

$$\text{MSE} = \frac{\text{Sum}(\text{Data} - \text{recover message})^2}{\text{Count}}$$

where;

Data - Generated random bits.
Recover message - Message recovered after decryption.
Count - Size of bits.

MSE = 0 denotes no error and successful recovery of original data. MSE ≠ 0 denotes an error and the original data was not recovered successfully.

Below section gives an overview of RLE algorithm with some examples.

RLE algorithm:

It is a lossless data compression form in which the repetitive bits are kept in a single value. For example, if input is PPPQQTT then it will be compressed as 3P2Q2T in output in which count of repetition of each distinct alphabet is recorded. Similarly, if the input is 0 1 1 0 then it will be taken

as 1 2 1 in output in which count of consecutive 1s (which equals 2) and 0s (which equals 1) is recorded. ‘First bit (which is 0) is mentioned as 1. Below Fig 1. Shows a block diagram of working of RLE algorithm.

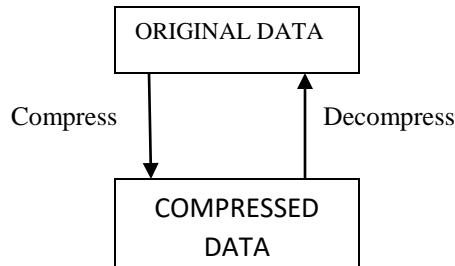


Fig.1 Working of RLE

RLE takes the original data and compresses it to get compressed data which in turn reduces the space complexity. To make it more secure, this paper has reduced the data storage problem by combining RLE algorithm with ECC.

RLE can also decompress the compressed data and can recover the original data.

II. LITERATURE REVIEW

ECC has proved to provide equivalent level of security with key of smaller sizes. The research in the field of ECC is mostly concentrated on its implementation on application specific systems. Such systems have confined resources like processing speed, storage, and domain specific CPU architecture.[6]. Moreover, in 2013, research work [8] has successfully extracted the security flaws and proposed an ECC-based scheme in addition to the password update and secured password authentication which is effectively aimed to guard against several related attacks. Additionally, different fault attacks for elliptic curve digital signature algorithm were proposed along with fault injection technique in ECC and the implementation of scalar multiplication to determine the secret signing key [10]. Recently research was also carried out for the implementation of doubling and point addition in Verilog system which is used in elliptic curve point multiplication for point addition and doubling, modular addition, modular squaring and then projecting to coordinate systems [5] (referred as base paper).

III. PROBLEM STATEMENT

In the existing system, there is no provision to calculate compression ratio and the total time taken for encryption of compressed data. Also, once the

original data gets recovered, there is no assurance whether the recovery of data was successful or were there any errors encountered.

IV. OBJECTIVES AND PROPOSED ALGORITHM

The main objectives of our proposed algorithm is to calculate the data compression ratio once the original data is compressed using RLE algorithm, calculate the time taken to encrypt the compressed data and finally calculate MSE to determine whether the original data was recovered successfully or not.

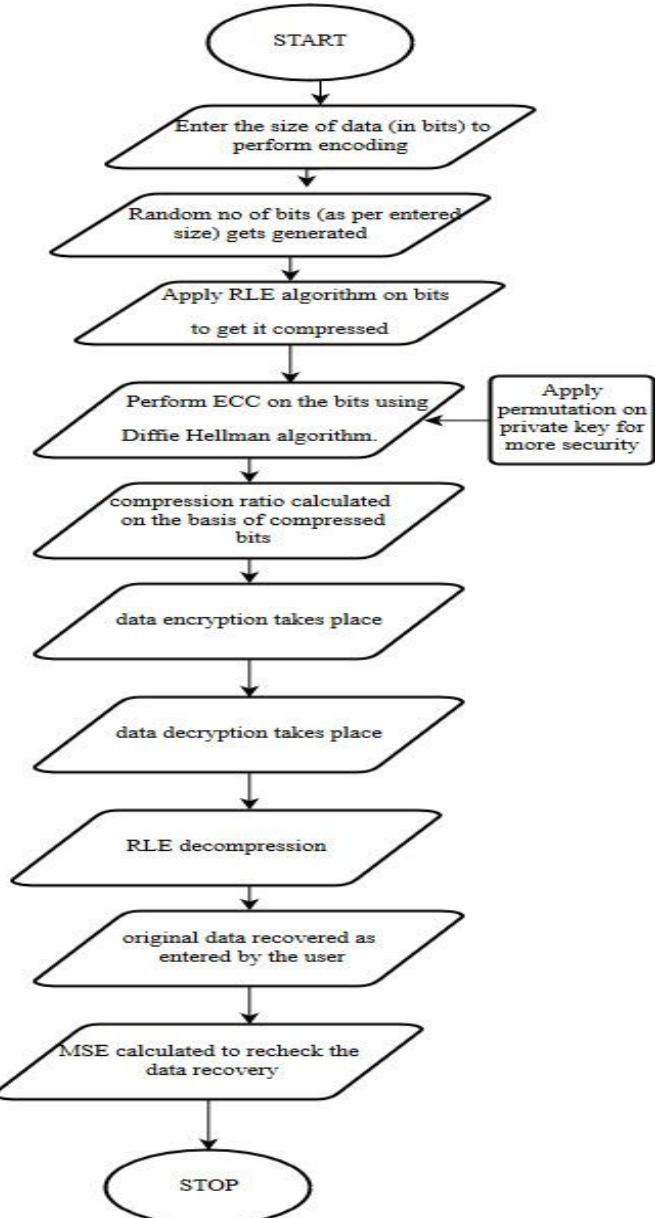


Fig.2. Complete flow of proposed system

Above Fig.2 is the flowchart which shows the complete working of our proposed algorithm. This flowchart overcomes the problems mentioned in Section III and achieves the objectives as mentioned in Section IV. Above calculations will not only help in reduction of time complexity but will also identify status of recovered original message. Proposed algorithm to achieve the above mentioned objectives is depicted. It shows the step wise execution of the process to achieve the objectives.

V. RESULTS

This section gives the results of the implementation work performed using MATLAB platform. Table 1 lists five different cases and gives a comparison of size of bits between the ECC and proposed algorithm. Last column calculates the data compression ratio by using the formula as mentioned in Section 1.

TABLE 1
DATA SIZE COMPRESSION RATIO

Case No	ECC	ECC+RLE Algorithm	COMPRESSION RATIO (IN %)
1	25	15	60
2	50	23	46
3	80	43	53
4	125	62	49
5	150	73	48

In the algorithm, data security is achieved by generating a new private key every time using permutation method (formula used in MATLAB-P=Perms (bint)).

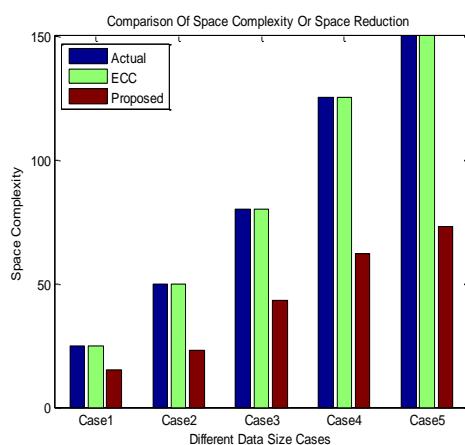


Fig.3. Data compression ratio graph

Fig.3 shows the graphical representation of the space complexity or space reduction between the existing ECC and the proposed system. This is the case wise

depiction of data according to TABLE 1. In Case 1, data size (in bits) was 25 and in the proposed algorithm it got reduced to 15. This helped in achieving the data compression ratio of 60%.

Below Table 2 gives a case wise comparison of time taken for encrypting data between the base paper work and proposed work.

TABLE 2
TIME TAKEN FOR ENCRYPTION

Case No	BITS	ECC (SEC)	ECC+RLE (SEC)
1	25	5.0721	2.8008
2	50	0.3066	0.2677
3	80	0.2123	0.2110
4	125	0.2464	0.2255
5	150	0.2639	0.2211

In Case 1 the time taken for data size of 25 bits in base paper is 5.0721 seconds which got reduced to 2.8008 seconds in new algorithm. This has helped in efficiently reducing the encryption time in proposed algorithm which has resulted in reduction of time complexity..

Fig.4 is the graphical representation of case wise reduction in time complexity for all the five cases as mentioned in Table 2.

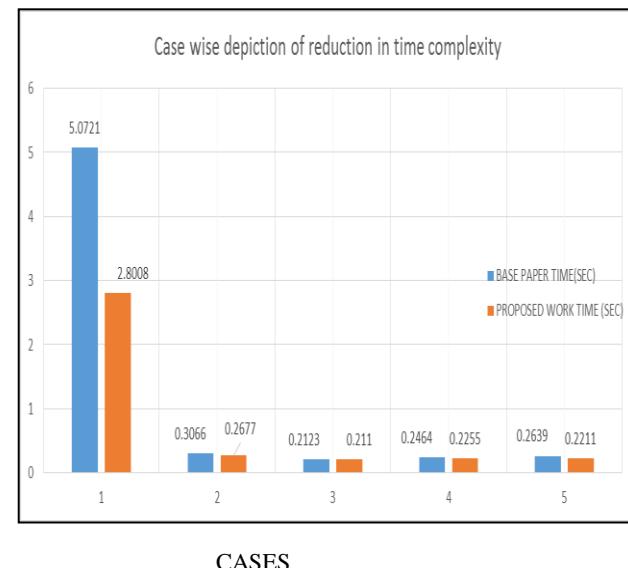


Fig.4 Case wise depiction of reduction in time complexity

VI. CONCLUSION

By using this algorithm we have shown the complete implementation of the system in which the algorithm calculated the data compression ratio to know the extent of compressed data and in turn this has helped to reduce the data storage complexity. Also it calculated the time taken for data encryption on different data bits and have plotted graphs of both time comparison and data compression ratio. After the execution of complete process, it calculates the mean square error to identify whether the data recovered successfully or not. These steps helps in making ECC algorithm more secure and efficient.

VII. LIMITATION AND FUTURE WORK

In the existing scenario data complexity or the data storage problem got reduced by using RLE algorithm and data security got increased by using permutation method for private key generation [2]. The proposed system calculates data compression ratio and time complexity for the input data which is in the form of bits – 0 or 1. In future, we can include input data in the form of alphabets, strings or sentences. This can make the ECC system more scalable and robust.

References

- [1]. Abdelhamid Tadmori, Abdelhakim Chillali, M'hammed Ziane, Cryptography over the elliptic curve (), Journal of Taibah University for Science, Volume 9, Issue 3, 2015, Pages 326-331, ISSN 1658-3655, <http://dx.doi.org/10.1016/j.jtusci.2015.02.005>.
- [2]. Ayushi Mathur, Dr. Varun Prakash Saxena, Data Compression And Security in Elliptic Curve Cryptography with Run Length Encoding ,International Journal of Computer Science and Networks, Volume 6, Issue 5, October 2017, Pages 575-579 , ISSN – 2277-5420.
- [3] Khalid Javeed, Xiaojun Wang, Mike Scott, High performance hardware support for elliptic curve cryptography over general prime field, Microprocessors and Microsystems, Volume 51, 2017, Pages 331-342,ISSN0141,9331,<http://dx.doi.org/10.1016/j.micpro.2016.12.005>.December 2016
- [4]. Lejla Batina, Siddika Berna Örs, Bart Preneel, and Joos Vandewalle. 2003. Hardware architectures for public key cryptography. *Integr. VLSI J.* 34, 1-2 (May 2003), 164. DOI=[http://dx.doi.org/10.1016/S0167-9260\(02\)00053-6](http://dx.doi.org/10.1016/S0167-9260(02)00053-6)
- [5]. M. M. Panchbhai and U. S. Ghodeswar, "Implementation of point addition & point doubling for Elliptic Curve,"*2015 International Conference on Communications and Signal Processing (ICCP)*, Melmaruvathur, 2015, pp. 0746-0749.doi: 10.1109/ICCP.2015.7322589", IEEE, 2015
- [6]. Rahat Afreen et al, "A Review On Elliptic Curve Cryptography For Embedded Systems", International Journal of Computer Science & Information Technology, Vol. 3, No. 3, Pp. 84-103, June 2011. DOI: 10.5121/ijcsit.2011.3307
- [7]. Ruchika Markan et al, "Literature Survey on Elliptic Curve Encryption Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, No. 9, Pp. 906-909, September 2013
- [8]. SK Hafizul Islam, G.P. Biswas, Design of improved password authentication and update scheme based on elliptic curve cryptography, Mathematical and Computer Modelling, Volume 57, Issue 11, 2013, Pages 2703-2717, ISSN 0895-7177, <http://dx.doi.org/10.1016/j.mcm.2011.07.001>.
- [9]. Sonali Nimborkar, Latesh Malik, Comparative Analysis of Authenticated Key Agreement Protocols Based on Elliptic Curve Cryptography, Procedia Computer Science, Volume 78, 2016, Pages 824-830, ISSN18770509,<http://dx.doi.org/10.1016/j.procs.2016.02.065>, 2016
- [10].Deepti Jyotiyan, Varun Prakash Saxena ,(ICT4SD 2016 GOA) "A Fault Attack for ScalarMultiplication in Elliptic Curve Digital Signature Algorithm. In: Vishwakarma H.,Akashe S.(eds) Computing and Network Sustainability. Lecture Notes in Networks and Systems, vol 12.Springer, Singapore,DOI: https://doi.org/10.1007/978-981-10-3935-5_29.IEEE, 2013
- [11]. Deepti Jyotiyan,Varun Prakash Saxena (Dec 23-25 2016);Fault attack for scalar multiplication over finite field ($E(F_q)$) on

- Elliptic Curve Digital Signature Algorithm," 2016International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur,2016,pp.14.DOI10.1109/ICRAIE.2016.7939539.
- [12] P. Nalwaya, V. P. Saxena and P. Nalwaya, A Cryptographic Approach Based on Integrating Running Key in Feedback Mode of ElGamal System,; 2014 International Conference on Computational Intelligence and Communication Networks, Bhopal, 2014,pp.719724.doi:10.1109/CICN.2014.157.
- [13].Anubhav Saxena ,Varun Prakash Saxena Sandeep Mal(April 2015) "Implementation of Fault Attacks on Elliptic Curve Cryptosystems" International Journal of Research in Advent Technology (IJRAT), Vol.3, No.4, April 2015 E-ISSN: 2321-, 9637.
- [14] PriyaNalwaya ,VarunPrakash Saxena(2014). "A Novel Cryptographic Approach Based On Feedback Mode Of Elgamal System". International Journal of Advance Research in Science & Engineering (IJARSE)- ISSN – 23198354.
- [15.] A. Barenghi, G. Bertoni, A. Palomba and R. Susella, "A novel fault attack against ECDSA," 2011 IEEE International Symposium on Hardware-Oriented Security and Trust, San DiegoCA,2011,pp.161-166.doi: 10.1109/HST.2011.5955015
- [16.]Hui Li, Ruixia Zhang, Junkai Yi, Hongqiang Lv,"A Novel Algorithm for Scalar Multiplication in ECDSA", 2012 Fourth International Conference on ComputationalandInformationSciences,vol.0 0,no.,pp.943946,2013,doi:10.1109/ICCIS.2013.254
- [17.] Ling, Jie & King, Brian. (2013). Smart card fault attacks on elliptic curve cryptography. Midwest Symposium on CircuitsandSystems.12551258.10.1109/MWSCAS.2013.6674882.
- [18].Rashidi, Bahram & Sayedi, S.M. & Rezaeian Farashahi, Reza. (2016). High-speed hardware architecture of scalar multiplication for binary elliptic curve cryptosystems. Microelectronics Journal. 52. 49-65.10.1016/j.mejo.2016.03.006.
- [19].Lavanya, M & Praveenkumar, G & Lena Murugan, N & Vigneshwaran, M & Saravanan, S. (2016). Authentication scheme for client and server using elliptic curve cryptography. International Journal of Pharmacy and Technology. 8. 25317-25325.
- [20].Mrabet, Amine & El-Mrabet, Nadia & Lashermes, Ronan & Rigaud, Jean-Baptiste & Bouallegue, Belgacem & Mesnager, Sihem & Machhout, Mohsen. (2017). High-Performance Elliptic Curve Cryptography by Using the CIOS Method for Modular Multiplication. 185-198. 10.1007/978-3-319-54876-0_15.
- [21].Phalakarn, Kittiphop & Phalakarn, Kittiphon & Suppakitpaisarn, Vorapong. (2016). Parallelized Side-Channel Attack Resisted Scalar Multiplication Using q-Based Addition-Subtraction k-Chains. 140-146. 10.1109/CANDAR.2016.0035.
- [22]. S. R. Singh, A. K. Khan and S. R. Singh, "Performance evaluation of RSA and Elliptic Curve Cryptography," 2016 2nd International Conference on Contemporary Computing and Informatics(IC3I),IEEE Noida,2016,pp.302306.doi:10.1109/IC3I.2016.7917979
- [23]. N. Alimi, Y. Lahbib, M. Machhout and R. Tourki, "On Elliptic Curve Cryptography implementations and evaluation," 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP),IEEE, Monastir, 2016, pp. 3540.doi:10.1109/ATSIP.2016.
- [24]. M. M. Chauhan, "An implemented of hybrid cryptography using elliptic curve cryptosystem (ECC) and MD5," 2016 International Conference on Inventive ComputationTechnologies(ICICT),Coimbatore,2016,IEEE,pp.16.doi: 10.1109/INVENTIVE.2016.7830092U
- [25]. S. R. Singh, A. K. Khan and T. S. Singh, "A critical review on Elliptic Curve Cryptography," 2016 International Conference on Automatic Control and Dynamic Optimization Techniques

(ICACDOT), Pune, 2016, pp. 13-18.
doi:10.1109/ICACDOT.2016.7877543URL:
<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7877543&isnumber=7877540>

- [26]. X. Fang and Y. Wu, "Investigation into the elliptic curve cryptography," 2017 3rd International Conference on Information Management (ICIM), Chengdu, 2017, pp. 412415. doi:10.1109/INFOMAN.2017.7950418 URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7950418&isnumber=7950330>
- [27]. M. Indaco, F. Lauri, A. Miele and P. Trotta, "An efficient many-core architecture for Elliptic Curve Cryptography security assessment," 2015 25th International Conference on Field Programmable Logic and Applications (FPL), London, 2015, pp. 1-6. doi: 10.1109/FPL.2015.7293950 URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7293950&isnumber=7293744>
- [28]. Kamal Kamal, Radu Muresan, "Capacitive physically unclonable function", Electrical and Computer Engineering (CCECE) 2017 IEEE 30th Canadian Conference on, pp. 1-6, 2017.
- [29]. Hamad Marzouqi, Mahmoud Al-Qutayri, and Khaled Salah. 2015. Review of Elliptic Curve Cryptography processor designs. *Microprocess. Microsyst.* 39, 2 (March 2015), 97-112.
DOI=<http://dx.doi.org/10.1016/j.micpro.2015.02.003>

ENHANCED MULTI-MODAL BIOMETRIC BASED SECURITY SCHEME WITH FEATURE BASED MACHINE LEARNING APPROACH

Rakhi Choudhary¹, Er. Krishan Kumar², Dr. Himanshu Monga³

M.Tech (Scholar), Assistant Professor, Principal

Department of Computer Science Engineering

Jan Nayak Chaudhary Devi Lal Memorial College of Engineering

Rakhichoudhary930@gmail.com¹, chhaperwal@gmail.com², himanshumonga@gmail.com³

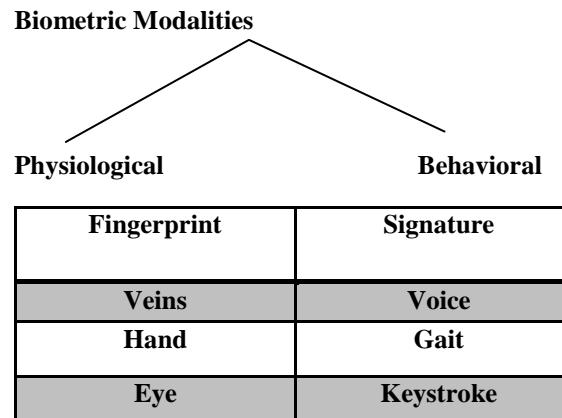
Abstract –The multi-modal biometric system, which defined that the behavioural and psychological features to verify a user. In real-time applications mostly used a uni-modal biometric system to security or authenticate user. An identification based on various biometrics represents a developing feature. It defined a multi-modal biometric system, which integrates face and speech identification in making user verification. Now a days, the data security and authentication system uni-modal user have become an important factor in-secure system. Normal Information like a passwords, key device such as a smart card, ID cards and numbers not reliable and improve techniques in secure environment. The main problem was what grade environments are to be removed and how the cost aspect could reduce, as the quality of structures up-surges the inconsistency of the intra particular illustrations due to delay increases in between repeated acquirements of the illustration also growth. The proposed model, discussed with the multi-modal biometric system integrating facial speech in making a personal documentation was introduced. The multi-modal biometric system in facial recognition and Speech recognition using ICA and GTCC algorithms features. The procedure information was combined with the help of fusion methods like score-level, feature level and sensor levels each system computers its own matching score. Implement the classification method like Convolutional Neural Network using Training and Testing Section. These were system with MATLAB 2016a Simulation tool to provide preferable false acceptance rate, false rejection rate, accuracy and precision enjoying efficient recognition.

Keywords- Multi-model Biometric; Authentication System; Data Security; GTCC; ICA and Score Level Fusion.

I. INTRODUCTION

Biometric System is used in a world-wide several of applications that require the verification and identification methods to confirm the verify of an individual[1]. In a real-time applications mostly used uni-modal biometrics to authenticate or recognize the human being or consumers. Uni-

modal biometric system arises some problems such as interference in sensed data, non-universality, multiple duplicacy and some variations etc. to resolve these limitations by participating various sources of data which is known as a multi-modal biometric system[2]. Multi-modal biometric systems association biometric information from various sources to establish the authenticity of a person. Verification in, multi-modal biometric system solve, to a degree, the problem postured by non universality. It is completed by taking into explanation various biometric traits that could better verify user when second-hand in conjunction as different to an individual modality[3]. This system also act as a preventive to duplicate hijackers by creating it more not easy to repeat the data because any illegal use would need to subject to imitate various features in face and speech like voice features and nose, lips and cheeks etc [4].



Most recently, developed biometric traits are unimodal system they rely on an individual unique feature to verify a person. Through, some extracted features like facial, iris, palm and fingerprint. Attaching the multiple modal biometric system could successful enhance the recognition rate of a system outside increase the population-area, define duplicate hijackers, optimizing the errors [5].

II. RELATED WORK

It gives the researcher a better understanding of the matter. Review of the literature in an early and essential step for conducting research. Only the research studies which are directly related to the

present study have been reviewed. A lot of research on speech and Face Fusion has been done in India and other countries in different age groups in which some important studies are given below:

In Biometric model [6] described that the new usage and efficiency of learning, all in all, and convolutional neural systems, a person, for programmed instead of hand-made component reflection for the hearty face acknowledgment crossways time pass. A CNNdevelopment utilizing the VGG-Face profound studyingare search to collect exceedingly discriminative and interoperable highlights that are beneficial to mature contrasts level over a blend of multi-model biometric datasets.Various [7] ID and check frameworks were currently best in class, be that as it may, their introductions stay inadmissible confronting to the expanding security needs. For the most part, the utilization of just a single biometric diminishes the dependability of these plans; hence, they need to affiliation a few modalities. It proposes a multi-biometric combination approach for personality confirmation utilizing two modalities: the fingerprints and the mark. Blends of neural multi-layer perceptron (MLP) are utilized for the unimodal characterization.[8]The projected technique applied on a artificial multi-modal biometrics authentication database. The last was shaped from Casia database and US-TB two datasets which define multi modal biometric image sets correspondingly.[9] Reviewed of the highlights, qualities, and limits of existing quality assessment system in unique finger impression, retina, and facial biometric were additionally possible. ultimately, a courier set of value measurements from these thrice modalities were assessed on a multi-modal database comprising of2D pictures, to value their execution with a concession to coordinate score got from the cutting edge acknowledgment frameworks[10].

III. MAIN ISSUES IN MULTI-MODAL BIOMETRIC SYSTEM

From the survey we have found the problem or future work in which we are going to continue our work of Fusion in Multimodal Biometric using speech and face [11].

Research Gap/Problem [12]:

Multi-modal biometrics system is the arrangement of double or further modalities like, iris, face, speech and ear modalities.The proposed model a face recognition system and speech ID system is collective as these functions are worldwide conventional and expected to harvest[16].Though the grouping of multi-modal improves safety and exactness, however the complication of the

structure growths due to improved numeral of landscapes removed of the several illustrations and smarts from further total in terms of purchase time[15].Now, this day a key problem is what grade structures are to be removed and how the cost aspect can be reduced, as the quantity of structures upsurges the inconsistency of the intra particular illustrations due to bigger delay in between repeated acquirements of the illustration also intensifications [13].Variety is increases of the organization will extra increase false acceptance rate (FAR). Accordingly, to determine this problem a current fusion-level and biometric fusion model is essential [14].The proposed work tiers to present a new user biometric authentication system based on a mutual acquisition of facial and language or voice with highly accurate rate, true positive and rejection rate.

IV. PROPOSED APPROACH

This research work includes a set of purposes that is associated with milestone of this process. The objectives are declared below. Now, we discussed the various existing techniques in speech recognition and facial recognition.The proposed model , the major aim is to present the recital of the interactive multi-modal biometric authentication system based-on user reliant on weighted synthesis approach.In facial detect or recognize , feature vector of the iris knowledge based data is resultant from component features and classified technique used. The feature vector is the speech model in the knowledge base. In Speech acknowledgment the characteristic component are combinations of still and self-motivated structures which has been removed by characteristic and classify the data through CNN algorithm. Feature extraction algorithm used for unique properties identifying, thus obtained is the face database in the information set. In the verification stage, the similar mark of the assessment framework and the preparation structure are consequent. Classify the both biometric model.

Steps in Feature Extraction Algorithm

1. **Center the data :** It observed all data \mathbf{Y} is calculated and the sum is subtracted from the considered data set to create it zero sum.

$$\mathbf{Yc} = \mathbf{Y} - \mathbf{F}\{\mathbf{Y}\}$$

2. **Cleaning the data :** Covariance Matrix Cov mat \mathbf{Y} is center of the data \mathbf{Yc} is calculated. The E decomposed of the cov \mathbf{Y} is evaluated. If DD is the E matrix and Eigen Values is the V vector of the eigen matrix then

$$\mathbf{X} = \mathbf{DD} - \frac{1}{2} \mathbf{E}(\mathbf{Yc})$$

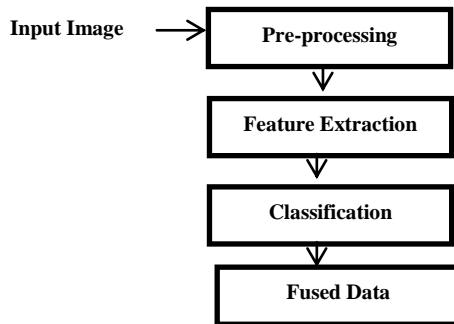


Fig 1. Proposed Model in Multimodel Biometric process

3. **Fix point repetition for one unit :** The speed is fast in ICA algorithm for 1 unit estimation and oneRow of the de-mixing matrix as a V i.e, processed iteratively.
4. **Calculate the binary independent components.**

V. RESULTS AND DISCUSSIONS

In this section, we described the multimodal biometric fusion with FACIAL and SPEECH Recognition in enhance the data security.

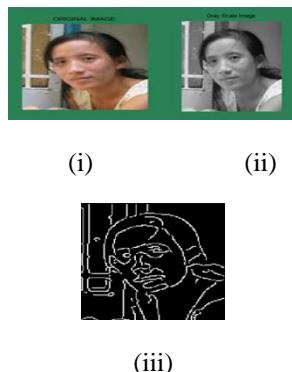


Fig 2. (i) Image Sample (ii) Gray scale Image and (iii) edge detection image (Canny)

The above figure shows that the upload the original image. To convert the original image to gray scale format. In Gray scale format cause of decrease the pixel. The edge detection using canny properties means calculate the maximum, minimum and average value.

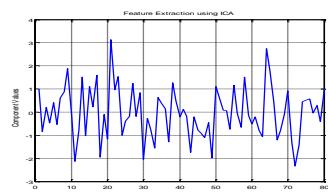


Fig 3. Feature Extraction with ICA

The above figure shows that the extracted features with ICA algorithm. It extract the unique properties

of the face image and component based feature extracted.

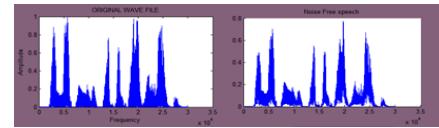
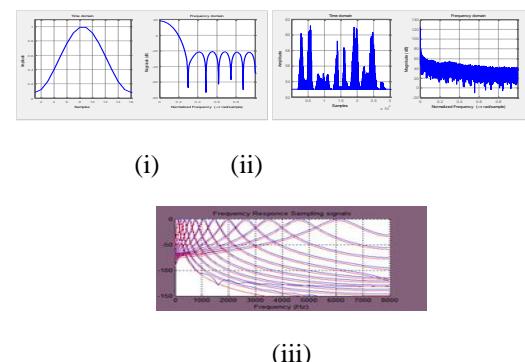


Fig 4. Sample Speech and Noise Free Signal

The above figure shows that the speech recognition module , first upload the wave form, feature extract using GTCC algorithm and fused the multimodel biometric with score level fusion method save in the data matrix form.



**Fig 5 (i) Time and Frequency Domain in Line
(ii) Time and Frequency Domain in Spectrum
and (iii) Frequency Response**

Above figure defines the time domain and frequency domain in spectrum format according to the amplitude and magnitude. Time-space alludes to a variety of sufficiency of the flag with time. So in recurrence space, over the whole day and age of recording, how often each pinnacle comes is recorded. Frameworks are dissected in the time area by utilizing convolution. Moreover, the DFT can be utilized to speak to each yield motion in a comparative shape. This implies any direct framework can be totally portrayed by how it changes the adequacy and period of cosine waves going through it.The Score level Fusion Apply for face and speech recognition using fused the data.

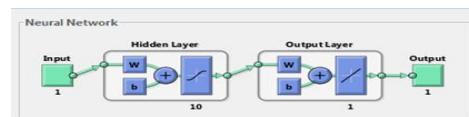


Fig 6. Convolutional Neural Network

This architecture represents the training processor in the form of neural network. We defined the iteration is 1000 for show the training performance, time and validation checks. This single-layer architecture used for classification purpose means first train the system through the algorithm and

validate the system to identify how accurately work through performance parameters.

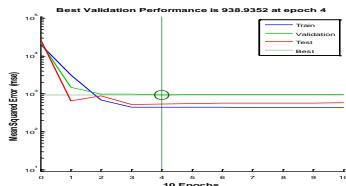


Fig 7. Best Performance

The figure defined that the best execution esteem is 938.25 at 4 ages concerning Mean Square Mistake rate. In this figure blue line demonstrates the preparation, which we offer up to 3 and green line demonstrates the approval of the framework execution and the red line demonstrates the testing on the framework.

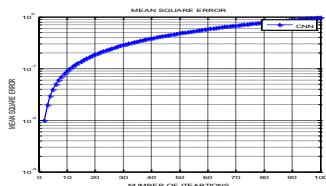


Fig 8. Means Square Error Rate

The figure defines; the mean square error rate (MSE) or mean squared deviation (MSD) means average of the training and testing module error. This is the important parameter which has found because find the average of error result. The average error value is identifies i.e 0.89.

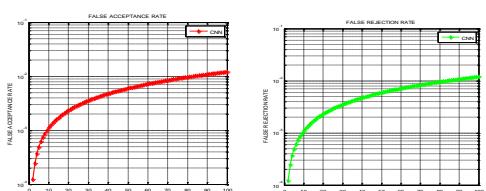


Fig 9. False Acceptance Rate and False Rejection rate

The figure appears; false acknowledgment rate implies positive information discover utilizing characterization in the testing Module and Concentrate the special Highlights. The false acknowledgment rate recognizes the esteem is the satisfactory blunder is 0.01889. Figure shows, the false rejection rate (FAR) means negative data collect using CNN for classification and feature identifies the scale invariant feature transform. The false rejection rate (FAR) compute the value is 0.0081.

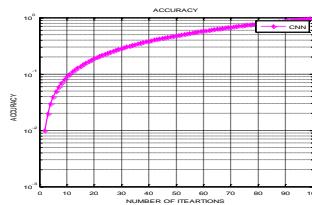


Fig 11. Accuracy in CNN

This figure defines; to compute the accuracy throughout the whole system. This is define the system has exact working. We identified the accuracy value is 99%.

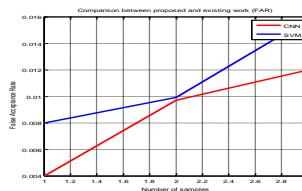


Fig 12. Comparison between FAR with proposed and existing work

The figure shows; comparison of the false acceptance rate means positive data find using classification in the testing Module and Extract the unique Features. The false acceptance rate identifies the value is the proposed acceptable error is 0.01889 and existing acceptance error is 0.98. The False Acknowledgment rate (FAR) is the likelihood that the framework mistakenly approves a non-approved individual, due to erroneously coordinating the biometric contribution with a layout.

Table 1. Comparison Between FAR with proposed and Existing work

No. of samples	FAR (Base)	FAR(proposed)
Face and Speech	0.0099	0.008
Speech and Face	0.0097	0.004

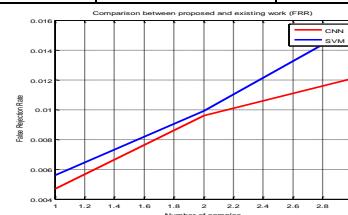


Fig 13. Comparison between FRR proposed and Existing work

Figure shows, comparison of the false rejection rate (FAR) means negative data collect using Feed forward neural network (FFNN) for classification and feature identifies the scale invariant feature

transform. The false rejection rate (FAR) compute the proposed value is 0.0081 and existing value is 0.0046.

Table 2. Comparison between FRR with proposed and existing work

No. of samples	FRR (Base)	FRR(proposed)
Face and Speech	0.0056	0.0047
Speech and Face	0.0099	0.0096

VI. CONCLUSION AND FUTURE SCOPE

In this conclusion , the proposed work check framework in view of facial and discourse. In the proposed framework another strategy is produced at score level combination to expand the execution of the face and discourse validation framework. In this right off the bat multimodal framework is produced utilizing ICA calculation and GTCC as it were. After that FAR, FRR and exactness have been assessed in which PCA performs well-having comes about like For ICA and CNN Precision = 97%, FAR= 0.01831, FRR= 0.00815. From the diagrams, it has been inferred that Autonomous segment examination and GTCC system functions admirably.Future works could go toward utilizing Hereditary calculation or ICA in hybridization with BFO. Free Segment Investigation (ICA) is a computational strategy to get concealed estimations of arbitrary factors. ICA essentially intended for multivariate information. The information utilized for examining utilizing ICA can be begun from many fields like financial aspects, computerized pictures, record databases and so forth. Additionally, Firefly advancement Calculation is all the more intense for the issues with a few measures of factors given.

REFERENCES

- [1] Snelick, Robert, Umut Uludag, Alan Mink, Mike Indovina, and Anil Jain. "Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, no. 3 (2005): 450-455.
- [2] Tuyls, Pim, Anton HM Akkermans, Tom AM Kevenaar, Geert-Jan Schrijen, Asker M. Bazen, and Raimond NJ Veldhuis. "Practical biometric authentication with template protection." In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pp. 436-446. Springer Berlin Heidelberg, 2005.
- [3] Wang, Jingyan, Yongping Li, Ping Liang, Guohui Zhang, and Xinyu Ao. "An effective multi-biometrics solution for embedded device." In *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, pp. 917-922. IEEE, 2009.
- [4] Jain, Anil K., Arun Ross, and Salil Prabhakar. "An introduction to biometric recognition." *IEEE Transactions on circuits and systems for video technology* 14, no. 1 (2004): 4-20.
- [5] Monwar, Md Maruf, and Marina L. Gavrilova. "Multimodal biometric system using rank-level fusion approach." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, no. 4 (2009): 867-878.
- [6] El Khiyari, Hachim, and Harry Wechsler. "Face Recognition across Time Lapse Using Convolutional Neural Networks." *Journal of Information Security* 7, no. 03 (2016): 141.
- [7] Dehache, Ismahene, and Labiba Souici-Meslati. "A multibiometric system for identity verification based on fingerprints and signatures." In *Complex Systems (ICCS), 2012 International Conference on*, pp. 1-5. IEEE, 2015.
- [8] Ghoualmi, Lamis, SalimChikhi, and AmerDraa. "A SIFT-Based Feature Level Fusion of Iris and Ear Biometrics." *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*.Springer International Publishing, 2015.102-112.
- [9] Bharadwaj, Samarth, MayankVatsa, and Richa Singh. "Biometric quality: a review of fingerprint, iris, and face." *EURASIP Journal on Image and Video Processing* 2014.1 (2014): 1-28.
- [10] Conti, Vincenzo, et al. "Fingerprint and Iris Based Authentication in Inter-cooperative Emerging e-Infrastructures." *Internet of Things and Inter-cooperative Computational Technologies for Collective Intelligence*.Springer Berlin Heidelberg, 2013.433-462.
- [11] Bakshi, Sambit, et al. "Score level fusion of SIFT and SURF for iris." *Devices, Circuits and Systems (ICDCS), 2012 International Conference on*.IEEE, 2012.
- [12] Fu, Bo, Jie Lin, and GuiduoDuan. "Analysis of multi-biometric encryption at feature-level fusion." *Intelligent Control and Automation (WCICA), 2012 10th World Congress on*.IEEE, 2012.
- [13] Dagher, Issam, and RabihNachar. "Face recognition using IPCA-ICA algorithm." *IEEE transactions on pattern analysis and machine intelligence* 28, no. 6 (2006): 996-1000.
- [14] Fox, Niall A., Ralph Gross, Jeffrey F. Cohn, and Richard B. Reilly. "Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts." *IEEE Transactions on multimedia* 9, no. 4 (2007): 701-714.
- [15] Mishra, Richa, and V. Pathak. "Human recognition using fusion of iris and ear data." *International Conference on Methods and Models in Computer Science*. 2009.
- [16] Kalka, Nathan D., JinyuZuo, Natalia A. Schmid, and BojanCukic. "Estimating and fusing quality factors for iris biometric images." *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40, no. 3 (2010): 509-524.

Evaluating similarity using near set theory for Plastic surgery face images

Sujata G. Bhele

Electronics Engg,

Priyadarshini College of Engg

Nagpur, India

sujata_bhele@yahoo.co.

Vijay H. Mankar

Electronics and Telecommunication Engg,

Government Polytechnic,

Nagpur, India.

vhmankar@gmail.com

Abstract— Popularity of plastic surgery is increasing as it has attracted attention from the research community. But the non linear effects due to plastic surgery is difficult to tackle by existing face recognition system. This paper emphasizes on extracting perceptually relevant information from groups of objects based on their descriptions. Object descriptions are represented by feature vectors containing probe function values similar to feature extraction in pattern classification. Near set approach is used to measure the similarity of images using tolerance class and performance parameters between two hetero faces. Five image features are used and their performance parameters are evaluated. Based on average values of parameters (nearness measure, Hausdorff distance and Hamming distance) facial similarity is proved. It is observed that nearness and Hausdorff distance are two significant parameters to justify the similarity of two face images. Hetero faces include plastic surgery images with Blepharoplasty, Rhinoplasty and Lip surgery images. From the results it is observed that near set with tolerance class perform better for Blepharoplasty, Rhinoplasty and lips images.

Keywords- Near set; Hausdorff distance; Hamming distance; distance measure; Blepharoplasty; Rhinoplasty.

I. INTRODUCTION

Plastic surgery is popularly used to enhance the facial appearance by correcting features and treating facial skin to get younger look. Plastic surgery methods are beneficial to patients who are suffering from various disorders due to excessive structural growth of facial features. These methods correct the facial feature providing a change in appearance. The popularity of plastic surgery is increasing due to reduction in cost and time required. Due to plastic surgery there is a large amount of change in texture and facial geometry and because of this there is a large variation between pre and post surgery images. Therefore matching the post surgery images with pre surgery images becomes a difficult task. Variations caused due to plastic surgery are long lasting and it cannot be changed. Due to this reason plastic surgery is now established as a new and challenging area. Plastic surgery is divided into two types Global and local plastic surgery. The local plastic surgery is done by the individual to correct defects, anomalies or improving skin texture. It is used to reshape and restructure

facial features. Local surgery leads to change in geometric distance between facial features but the overall texture and appearance may look similar to the original face. Global surgery is used to entirely change the appearance, texture and facial features of an individual. The features of an individual are reconstructed to resemble the original face but it usually not the same. Because of this large change in facial features it becomes difficult to recognize the pre and post surgery face. This paper focuses on the following types of plastic surgery such as –

Rhinoplasty (nose surgery):- It is use to reconstruct the nose which may be damaged due to accident or involving birth defects. It also cure breathing problems caused due to nasal structure.

Blepharoplasty (eyelid surgery):- It may be used to reshape both upper as well as lower eyelid in cases where excessive skin tissues growth may cause vision problem.

Lip augmentation:-It involves proper shaping and enhancement of lips with injectable filler substances.

These surgeries change the appearance of face which leads to reduced face recognition performance. The pose, expression and illumination can be corrected possibly by requiring the user to repeat the enrolment procedure, plastic surgery is more like ageing where a repeated data capture operation cannot be expected to enhance biometric performance. Thus the plastic surgery of a face can change the appearance of face and texture to such an extent that it becomes difficult to predict which features are changed with unavailable surgery information. This paper considers the local surgeries such as Rhinoplasty, Blepharoplasty and lip surgery in which the facial components such as eyelid, nose and lip can be reshaped or restructured by plastic surgery. The local skin texture around the face component may also be disturbed. This problem can be handled if facial feature is represented as an object with description of features. This is presented in near set theory. This paper uses near set theory to find the similarity between the pre and post surgery images using feature values. The near term is used to find that pre surgery image is partially or completely matches the description of post surgery image. Near set has the advantage of reducing the search space and it will also increase the accuracy and speed. Near set uses statistical features to describe the face such as average gray value, normalized R, G, B values and entropy. Entropy is used to deal with uncertain variations in the face. Hamming

distance has the ability to calculate the difference between two sets or elements.

A. Related work

Face recognition algorithms use facial information by extracting features and process them. It is popularly studied using several approaches to address the challenges of illumination, pose and expression [1][2][3][4] and challenges of aging and disguise is addressed by [5][6], but the popularity of plastic surgery is increasing as it introduces new challenges in designing face recognition system. Plastic surgery changes the geometry and texture of face image. Therefore matching post surgery images with pre surgery images becomes a difficult task. The plastic surgery basically is used to improve facial appearance as reported in [7][8]. Plastic surgery is also used for removing scares, birthmarks and correcting disfiguring defects as reported in [9][10][11]. Singh et. al. proposed a method which analyzed and experimentally evaluated plastic surgery database images using various face recognition algorithms and concluded that plastic surgery images are difficult to identify using existing algorithms [12]. De Marsico et. al. presented an approach to integrate information obtained from local regions for matching pre and post surgery face images [13]. A sparse representation approach on local facial fragments proposed by Aggarwal et. al. to match surgery face images [14]. A multiobjective evolutionary granular algorithm is proposed in [15] which extract the granular information from the face at multiple levels. A face recognition method for plastic surgery faces is proposed in which features are extracted using shape local binary texture (SLBT) and periocular features and then cascaded for invariant recognition of plastic surgery faces [16]. Based on the experimentation carried out by authors it can be concluded that face recognition algorithms such as PCA, FDA, GF, LLA, LBP and GNN have shown recognition rate not more than 40% for local plastic surgery. This paper introduces perceptual resemblance of plastic surgery faces using near set. The basic of near set and the relation between the neighborhood and the tolerance classes is calculated by using nearness relation is presented in [17][18]. Near set theory is used in image correspondence [19] and segmentation evaluation in [20]. The recognition of images is done by extracting the texture pattern such as contour and finding the similarity using near sets as reported in [21]. Generalization of rough set theory gives near set .Suppose one set X is near to another set Y to the extent that the description of at least one of the objects in X matches with the description of at least one of the objects in Y. Rough set was initially introduced by Pawlak et. al. gives an idea of perception of objects based on the level of classes[22][23]. Due to plastic surgery the appearance of face changes greatly. The change in appearance of face affects the accuracy of face recognition methods reported by Zhao et. al. [24].This poses a new challenge in recognition of faces. Plastic surgery of faces changes the appearance of faces to such an extent that it becomes difficult to identify the facial change. This paper provides a framework for measuring the resemblance between pre and post plastic

surgery images using nearness measure, Hausdorff distance measure and Hamming distance measure.

II. PROPOSED METHODOLOGY

The similarity of pre and post surgery images are calculated using the steps shown in Fig. 1

A. Near Set Theory

In the proposed work, theory approach had been used to identify similar faces irrespective of surgery on any of the feature of the face including nose, mouth, eyelids and lips. The facial image similarity is based on nearness measure which depends on the degree of near sets for what amount they resemble to each other. It represents systematic approach to determine the degree of similarity between a pair of disjoint sets. The nearness measure was first proposed based on indiscernibility relation and equivalence classes. It includes cardinalities of the equivalence classes that have some descriptions for two sets to be more similar. There are more pairs of equivalence classes (matching features).

The nearness measure consequently is determined by number of objects in equivalence classes that have matching descriptions. The same approach can be generalized to tolerance classes.

The system considers five features of color images. These features are average gray scale value, normalized G, normalized R, normalized B and entropy.

To be more precise the image is divided into sub blocks of size 25x25 and the images are resized to 225x225 for equal dimension blocks. We evaluated our system for low dimension blocks including 5x5, 15x15 but the results does not showed much variations. For low computational complexity 25x25 block size was chosen.

The features of pre and post surgery images are calculated by considering image1 and image2 as follows:-

Average gray value: -The 25x25 color block was converted to gray scale and the mean value was taken. (AG1 and AG2)

Normalized R, G, B :- Let S1, S2 and S3 represents 25x25 block of R, G and B color space then for image 1,

$$D1 = \sum_n \sum S_1 + \sum_n \sum S_2 + \sum_n \sum S_3 \quad (1)$$

Similarly, for image 2,

$$D2 = \sum_n \sum S_1 + \sum_n \sum S_2 + \sum_n \sum S_3 \quad (2)$$

D1 and D2 represent the sum of all blocks of R, G and B color space.

The normalized value of R, G and B color space is calculated as given in eq. 3, 4, 5, 6, 7 and 8 for image 1 and image2.

Normalized R

$$N_{R1} = \sum_n \sum \frac{S_1}{D_1} \quad \text{for image 1} \quad (3)$$

$$\text{And } N_{R2} = \sum_n \sum \frac{S_1}{D_2} \quad \text{for image 2} \quad (4)$$

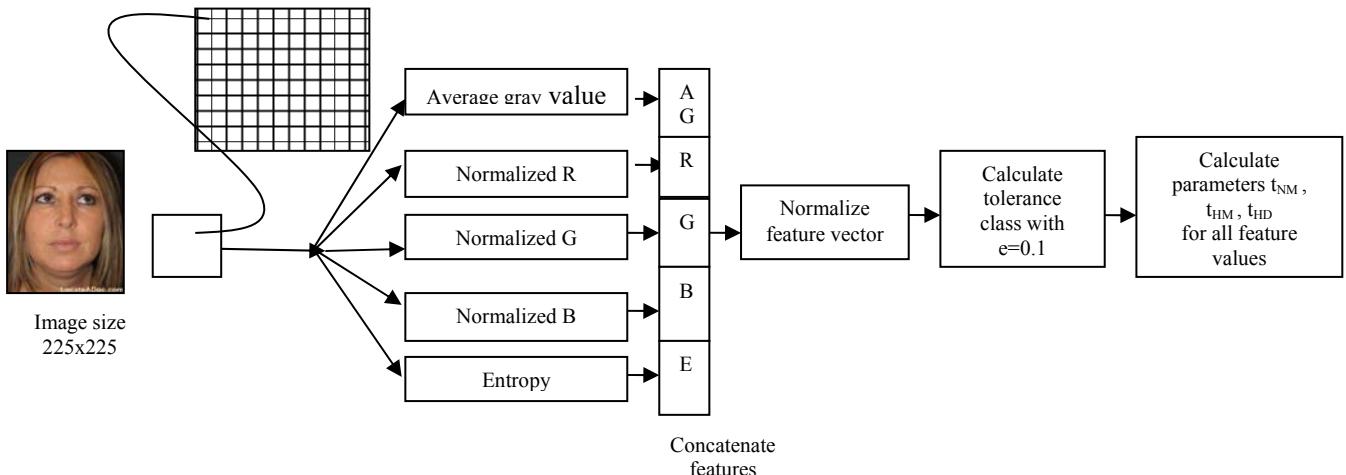


Figure 1 Block diagram of the proposed method using near set theory

Normalized G

$$N_{G1} = \sum_n \sum \frac{S_2}{D_2} \quad \text{for image 1} \quad (5)$$

$$N_{G2} = \sum_n \sum \frac{S_2}{D_2} \quad \text{for image 2} \quad (6)$$

Normalized B

$$N_{B1} = \sum_n \sum \frac{S_3}{D_1} \quad \text{for image 1} \quad (7)$$

$$N_{B2} = \sum_n \sum \frac{S_3}{D_2} \quad \text{for image 2} \quad (8)$$

Shannon Entropy

It is the information content, a high probability of occurrence specifies low information and low probability gives more information. The context can be applied to facial image for extraction of features. The factor gives the information contents where there are variations on faces. Shannon's definition of entropy suffers from following limitations such as it is undefined when p_i and a better measure of ignorance is $1-e_1$ rather than $1/e_1$. In this case the mathematical new definition of entropy is

$$E = \sum [e_1 * e(1 - e_1/255)] \quad (9)$$

Where e_1 is a row vector of gray scale image of block size 25x25

$$E_1 = Gray(I_{25x25}) \quad (10)$$

Further $e_1 = [e_1(:)']'$

The entropy is calculated for both the images.

For 225x225 image, number of block features are 81. Considering all five features for one image, feature matrix F_x is of size 81x5 and corresponding for image 2, F_y is 81x5 features for both images are concatenated as

$f = [F_x; F_y]$ for normalization

So f will have dimensions 162x5. Normalization is done by dividing all values in column by maximum value in that column further the features of image 1 and 2 are again separated in f_{xn} and f_{yn} matrix.

Now for the five features in columns, tolerance class is obtained for each column for tolerance value of 0.1.

a) Tolerance class ratio (Tc ratio)

Tc ratio is obtained as

If $n_x > n_y$

$$\text{Tc ratio} = O_{inY} / O_{inX}$$

Else

$$\text{Tc ratio} = O_{inX} / O_{inY}$$

Were S_x =number of features of image=81

Objects=25x25=225

$$n_x = \text{length}(\text{find}(c \leq S_x)); \quad (11)$$

Were c is row vector of tolerance matrix. Similar way n_y is calculated for image 2 with

$$n_y = \text{length}(\text{find}(c > S_x)); \quad (12)$$

Parameter calculations:-

The final table produces Nx4 vector which includes

b) Tolerance class size

c) Elements in class 1 (Image 1) [O_{inX}]

d) Elements in class 2 (Image 2) [O_{inY}]

We have T_1 , T_2 , T_3 , T_4 and T_5 with four columns corresponding to tolerance class size, objects in X (Image1/class 1), objects in Y (Image 2/class 2) and Tc ratio.

The row of T_1-T_5 varies depending on the similarity between two classes.

i. **Nearness measure**:- Let X and Y be two disjoint sets and let $Z=XUY$ then resemblance between two classes is calculated as,

$$t_{NM} = \text{sum}[T_1(:,1) * T_1(:,4')]/\text{sum}[T_1(:,1)] \quad (13)$$

ii. **Hamming measure**:- It was not meant for sets, hence it is modified. The idea behind estimating hamming measure is that it produces high values for classes which have objects in X that are close to objects in Y.

It is calculated as $t = \text{sum}(T_x)$; column sum of all four columns

Here t will have four values, sum of all four columns

$$t_{NM} = |t(1,2) - t(1,3)|/t(1,1) \quad (14)$$

Were $t(1,1)$ represents objects in X and Y both

$t(1,2)$ —Objects in X

$t(1,3)$ – Objects in Y

The idea behind this measure is that, for similar sets, the average feature vector of the portion of a tolerance class ($Z=X \cup Y$) that lie in X should have values similar to that lie in Y. This measure performs best with proper selection of tolerance value .

iii. **Hausdorff Distance**: - It is used to measure the distance between sets in metric space. This distance is calculated from each element in class 1 to every element of class 2. The shortest distance is taken as the infimum. The process is repeated for every $x_i \in X$ and the largest distance (supremum) is selected as the Hausdorff distance from X to Y. The same is repeated from set Y to set X since the distance is not necessarily same.

Before calculating the distance objects in X and objects in Y are divided by total objects in X and Y.

$$\text{Finally } t_{HD} = 1 - HD(X, Y) \quad (15)$$

Were HD is Hausdorff distance between X and Y.

Low value of HD corresponds to higher resemblance and vice versa. The performance of HD also depends on value of tolerance. It is poor for low values of tolerance. When tolerance value=0, tolerance class becomes equivalence class and HD=0, even for dissimilar images and the measure will produce a value 1.

III. EXPERIMENTAL SET UP

The dataset consists of 98 images having eyelid surgery, 14 having lips surgery and 164 having nose surgery. Therefore total dataset consists of $98x2+14x2+164x2=538$ facial images including before and after images for Blepharoplasty, lips augmentation and Rhinoplasty. The dataset is downloaded from American society for aesthetic plastic surgery 2008. Also some facial surgery images were acquired from local research institutes. In the real world it is difficult to identify the person undergone plastic surgery. Therefore face recognition algorithms should be designed robust to variations introduced by plastic surgery even in general operating conditions. In this paper we have considered images based on three surgeries such as Rhinoplasty, Blepharoplasty and lip augmentation as shown in Fig. 2. The following table shows the dataset made available in Table 1.The images from the database are resized to 225x225.

TABLE 1: PLASTIC SURGERY DATABASE

Plastic Surgery Procedure	Number of individuals
Nose Surgery (Rhinoplasty)	71
Eye-Lid-Lift surgery (blepharoplasty)	67
Others(Lip augmentation)	44



Figure. 2 The first two rows are the pre and post eyelid surgery(Blepharoplasty) , third and fourth row indicates pre and post nose surgery images(Rhinoplasty) and fifth and sixth row indicates pre and post lips surgery images

IV. RESULTS AND DISCUSSION

Experimental evaluation is done using the near set approach to prove the effectiveness of this method. The experiments are performed on plastic surgery database considering pre and post surgery images of Rhinoplasty, Blepharoplasty and Lip augmentation. 538 pre and post surgery images from the plastic surgery face database are used in this experiment. A single pre surgery image was evaluated for different post surgery images including its own post surgery image. Some of the comparative images are shown in Fig. 3 and 4 with the target pre surgery image.

The results obtained for some of the pre and post surgery images are described as follows:-

The Table 2, 3 and 4 shows that nearness measure and Hausdorff distance distinguishes its own post surgery image effectively related to other individuals post surgery images. The nearness measure value and Hausdorff distance value is high for the images of eyelid, lips and nose surgery images shown in Fig. 4, 5 and 6 respectively. The row values in the table are average values of nearness measure, Hausdorff distance and Hamming distance for all 5 features described earlier. From the average values it can be concluded that about 85-95% resemblance is shown by Hausdorff distance and about 70-80% resemblance is shown by nearness measure which is better as compared to other methods as shown in Table 5.

V. CONCLUSION

Plastic surgery procedures change the facial regions both locally and globally by changing the facial appearance, thereby posing a serious challenge to face recognition system. This paper presents an approach based on near sets along with two additional distance measures such as Hausdorff and Hamming distance to tackle the pre and post plastic surgery faces. We present an experimental study that evaluate the performance of proposed method on plastic surgery database that contain face images with local surgeries such as Rhinoplasty, Blepharoplasty and lip augmentation. The proposed method analyze the pre and post plastic surgery images using near set theory by considering five features of

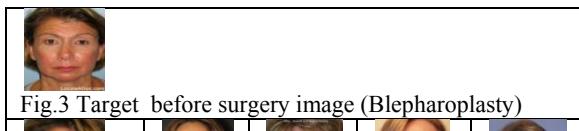


Fig.3 Target before surgery image (Blepharoplasty)



Figure. 4 Sample of after surgery images to be compared (Blepharoplasty)



Figure.5 Sample of Before and after lips surgery images.



Figure.6 Sample of Before and after Rhinoplasty (nose) surgery images.

TABLE2. PARAMETER VALUES FOR BLEPHAROPLASTY (EYELID) SURGERY IMAGES SHOWN IN FIG.4 AND 5.

Before Image	After Image	'tNM'	'tHD'	'tHM'
Image 1	Image 1	0.89306	0.92633	0.023912
Image 2	Image 2	0.761964	0.917331	0.054587
Image 3	Image 3	0.785471	0.877646	0.015021
Image4	Image4	0.770374	0.894428	0.015244
Image 5	Image 5	0.865242	0.91953	0.02923

TABLE 3. PARAMETER VALUES FOR LIPS SURGERY IMAGES SHOWN IN FIG.6.

Before Image	After Image	'tNM'	'tHD'	'tHM'
Image 1	Image 1	0.785893	0.854	0.032458
Image 2	Image 2	0.78334	0.865396	0.033472
Image 3	Image 3	0.716695	0.85715	0.05966
Image4	Image4	0.771476	0.872807	0.031966
Image 5	Image 5	0.903753	0.854321	0.017843

TABLE 4. PARAMETER VALUES FOR RHINOPLASTY (NOSE) SURGERY IMAGES SHOWN IN FIG.7.

Before Image	After Image	'tNM'	'tHD'	'tHM'
Image 1	Image 1	0.884254	0.954351	0.008059
Image 2	Image 2	0.817803	0.928873	0.024367
Image 3	Image 3	0.872721	0.9027456	0.043301
Image4	Image4	0.766184	0.888781	0.026188
Image 5	Image 5	0.838465	0.904974	0.039549

color images such as average gray scale value, normalized G, normalized R, normalized B and entropy. The distance is calculated by three distance measures such as Hamming, Hausdorff and nearness. The average values are calculated for Blepharoplasty, Rhinoplasty and lip surgery images. From the

TABLE 5. Algorithms used for performing plastic surgery face recognition rank-1 accuracies as listed below.

Authors	Algorithm used	Rank one accuracy
Singh et. al.	PCA FDA LFA CLBP SURF GNN	29.1% 32.5% 38.6% 47.8% 50.9% 54.2%
Marsico et. al.	Correlation based approach	70.6%
Bhatt et. al.	Evolutionary granular approach	78.6%
Aggarwal et. al.	Combination of recognition by parts and sparse representation approaches	77.9%

average values it can be concluded that nearness measure and Hausdorff distance distinguishes efficiently pre and post surgery images. The limitation of near set theory is that it works well for local surgeries as compared to global surgeries. In future the similarity may be improved by considering both shape and texture features.

ACKNOWLEDGMENT

Thanks to the reviewers for their attention to detail and many valuable suggestions.

REFERENCES

- [1] S. Li, R. Chu, S. Liao and L. Zhang, "Illumination invariant face recognition using near infrared images", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, vol. 29, no. 4, pp. 627-639.
- [2] R. Singh, M. Vatsa and A. Noore, "Improving verification accuracy by synthesis of locally enhanced biometric images and deformable model", Signal Processing, 2007, vol. 87, no. 11, pp. 2746-2764.
- [3] V. Blanz, S. Romdhani and T. Vetter, "Face identification across different poses and illumination with a 3D morphable model", 2002, Proceeding s of International conference on Automatic Face and Gesture Recognition, pp. 202-207.
- [4] Bhele S. G. and V. H. Mankar, "Recognition of faces using discriminative features of LBP and HOG descriptor in varying environment", Proceedings of the International Conference on Computational Intelligence and Communication networks, Dec 12-14 2015 , IEEE Explore press, India pp. 426-432.
- [5] N. Ramanathan and R. Chellappa, "Face verification across age progression,. IEEE Transactions on Image Processing, 2006, vol.15,no. 11, pp. 3349-3362.
- [6] N. Ramanathan, A.R. Chowdhury and R. Chellappa, "Facial similarity across age, disguise, illumination and pose",

Proceedings of International Conference on Image Processing, 2004, vol. 3, pp. 1999-2002.

[7] Poonam. Sharma, Ram.N. Yadav, Karmveer.V. Arya, “Pose-invariant face recognition using curvelet neural network”, IET Biometrics, 2014, Volume 3, Issue 3, p. 128 – 138, Sep.

[8] Sang.-Heon. Lee, Dong.-Ju. Kim, Jin.-Ho. Cho, “Illumination-robust face recognition system based on differential components”, IEEE Transactions on Consumer Electronics, 58 (3), Aug 2012, pp. 963–970

[9] Wilman.W.W. Zou, C. Pong Yuen, “Very Low Resolution Face Recognition Problem”, IEEE Transactions on Image Processing, 21 (1) Jan 2012, pp. 327–340

[10] Unsang. Park, Yiyi. Tong, Anil.K. Jain, “Age-Invariant Face Recognition”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32 (5) (May 2010), pp. 947–954

[11] Sivaram Prasad Mudunuri, Soma Biswas, Low Resolution “Face Recognition across Variations in Pose and Illumination”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 38 (5) (May 2016), pp. 1034–1040

[12] R. Singh, M. Vatsa, H. S. Bhatt, S. Bharadwaj, A. Noore, and S. S. Nooreyedan, “Plastic surgery: A new dimension to face recognition”, IEEE Trans. Inf. Forensics Security, 2010, vol. 5, no. 3, pp. 441–448, Sep.

[13] M. De Marsico, M. Nappi, D. Riccio and H. Wechsler, “Robust face recognition after plastic surgery using local region analysis”, Proceedings of International Conference Image Analysis and Recognition, 2011, vol. 6754, pp. 191–200. ISBN:978-3-642-21595-7

[14] G. Aggarwal, S. Biswas, P. J. Flynn and K. W. Bowyer, “A sparse representation approach to face matching across plastic surgery”, Proceedings Workshop on the Applications of Computer vision, 2012, pp. 1-7.

[15] Himanshu S. Bhatt, Samarth Bharadwaj, Richa Singh, Mayank Vatsa, “Recognizing Surgically Altered Face Images Using Multiobjective Evolutionary Algorithm”, IEEE Transactions on Information Forensics and Security, vol. 8, no. 1, January 2013.

[16] N. S. LakshmiPrabha, S. Majumder, “Face Recognition System Invariant to Plastic Surgery”, 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012.

[17] J.F. Peters, “Near sets. General theory about nearness of objects”, Applied Mathematical Sciences, 2007, 1 no. 53, 2609-2629.

[18] C. Henry, “Near Sets: Theory and Applications”, Ph.D. Diss., supervisor: J.F. Peters, Dept. Elec. & Comp. Engg., U. of Manitoba, WPG, MB, Canada, 2010.

[19] J. F. Peters, “Tolerance near sets and image correspondence”, International Journal of Bio inspired computation, 2009, vol. 1, no. 4, pp 239-245.

[20] C. Henry and J. F. Peters, “Near set index in an objective image segmentation evaluation framework”, Proceedings of the Geographic Object based Image Analysis: Pixels, Objects, Intelligence, University of Calgary, Alberta, 2010, pp. 1-8.

[21] Christopher Henry and James F. Peters, “Image Pattern Recognition using Near Sets”, Proceedings of Eleventh International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC 2007), Joint Rough Set Symposium (JRS 2007)

[22] Pawlak, Z.: “Rough Sets”, Institute for Computer Science, Polish Academy of Sciences, Report 431 (March 1981). Plenum Publishing Corporation

[23] Pawlak, Z.: “Classification of Objects by Means of Attributes”, Institute for Computer Science, Polish Academy of Sciences, Report 429 (March 1981).

[24] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, “Face recognition: A literature survey”, ACM Computing Surveys, Vol. 35, No. 4, pp. 399–458,

Face Image Recognition Based on Linear Discernment Analysis and Cuckoo Search Optimization with SVM

By

Jamal Mustafa AL-Tuwaijari

Suhad Ibrahim Mohammed

Department of Computer Science - College of Science - University of Diyala - Iraq

Abstract

Face image recognition became an effective research area over last two decades, It covers a wide range of activities from many aspects of life such as authentication and identification, airport security, inmate tracking, e-commerce and facebook's automatic tag. The aim of face image recognition is to recognize the face of a persons depend on the features extracted from their faces. In this paper, two proposed systems were developed, the conventional proposed system of image recognize include many steps to recognize faces. The first step is the preprocessing of images for all training and testing images. The second step is detecting accurate the accuracy of the face by using Viola and Jones algorithm. The third step is features extraction and selection by using linear discernment analysis (LDA). In the final step the support vector machine (SVM) is applied to reorganize the faces as known or unknown face. The proposed system has been implemented by using a datasets set (MUCT). This dataset is considered taking the processing of faces for frontal position. The results show that the SVM classifier recognition provides an accuracy rate of 99.25% with cuckoo search algorithm, and 96% without cuckoo search algorithm for the same test images.

Keywords—Face Detection ; Face Recognition ; Linear Discernment Analysis (LDA); Support Vector Machine (SVM) ; cuckoo search algorithm (CS); Feature Extraction ; MUCT. Viola-Jones

1. Introduction

Recently, face image recognition is a rapidly increasing field for its several uses in the several applications such as security, biometric authentication and neuromas other area. There are numerous problems that appear because to the exactness of several factors that affects the feature of image. When processing images one must take into account the variations in light, image quality, the persons pose and facial expressions along with others. The face image recognition is an essential ability of human, but it is hard for face image recognition systems to perform as well as human under different conditions, including illumination, variation of poses, expressions, occlusion..etc [1].

The face image recognition manly consists of four steps. The first step is the face detection which finds the interest area in the image that contains the face. The second step is the face extraction features which positions the face detected into an estimate pose, usually represented by a target face or model. The third step is face representation describes the face with certain aspects of interest, the final step is face classification which decides whether the representation belongs to a model or target face or not [2].

The detection phase is the first phase; it consist of identifying and locating a face in an image. The recognition phase is the second phase; it consist of feature extraction, where significant data for recognition is stored, and the matching, where the recognition result has been given with the help of a face database. Face classification has been an in process research area, and it must be used in vast range of applications. It is about identifying a person from one or every images of his/her face [3]. Feature extraction are to extract feature reduction method will be applied after face detection by using LDA. It is achieved by projecting the image onto the Eigenface space by LDA ,and used feature selection of optimization cuckoo search algorithm (CS) to enhancement the search then reduce size and increase speed rate recognition , finally used the result cuckoo search(CS) with support vector machine (SVM) method is widely used to classification in pattern recognition.

2. Methodology of Face Recognition System

2.1. Voila-Joins detection

Viola Jones image detection suggested by Paul Viola and Michael Jones in 2001 was one of the first methods to supply object with detection at very fast rates [4]. Viola and Jones method was adopted because it characterized by fast processing and high accuracy by applying robust algorithm and used accurate cropping of a face, eye, mouth, and nose regions from a detected image. It is the method for fast and to make a correction for object detection through AdaBoost machine learning [5].

2.1.1 Adaboost machine discovering based method

This method attempts to discover a particular Haar features in terms of the face of the human. This method has three meanings which are explained in the following [6].

- Integral Image: Here are the calculation values in pixels of the present image. The value at any location (x, y) in the integral image is the summary of the values of the image pixel upper and left side of position (x, y) defined as in equation (1):

$$ii(x, y) = \sum i(x', y') \quad \dots\dots\dots (1)$$

$$x' \leq x, y' \leq y$$

Where $ii(x, y)$ is the integral image and $i(x, y)$ is the original image.

- Haar features: We can calculate the results of any Haar feature when we multiply weights by calculated region of any individual rectangle. A Haar feature classifier computes the value of a feature using the integral of rectangular image. Several Haar feature classifiers compose a stage [7].
- Cascade Classifier: Calculation completely removes face postulants quickly using a cascade of stages. The cascade removes postulants by making exacting requirements in each stage with former stages will be much more difficult for a postulant to pass. Postulants exit the cascade if they pass all stages or fail any stage. A face is detected if a candidate passes all stages. This process is shown in Figure (1). Where T and F are the abbreviation of True and False respectively[8].

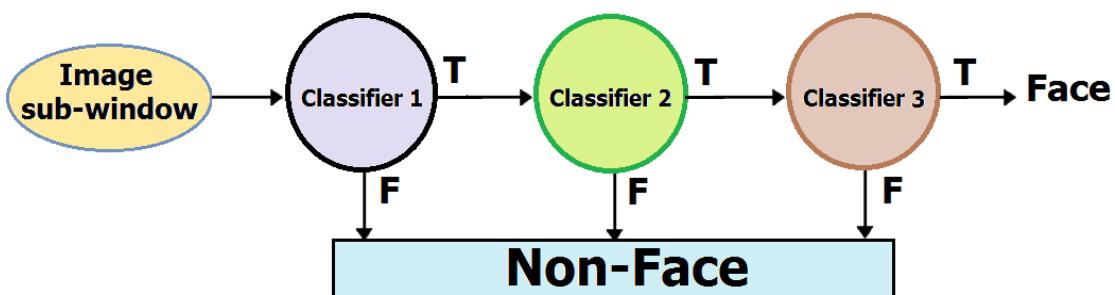


Figure (1) Cascade of Stages

2.2 Feature Extraction by using LDA

Linear discriminant analysis (LDA) or Fisher's linear discriminant (FLD) method is a considerably used for object classifications (faces) based on the extracted character from face image. LDA can be used as a dimension decreasing technique, which is used for classification purpose. The objective of LDA method is to get a low-dimensional character representation where the data can be classified according to their class labels. It can be used as a classification approach. LDA tries to define the best set of the discriminant projection vectors to map the original character space onto a lower-dimensional character space, by maximizing the Fisher criterion. LDA maximizes the ratio of between-class scatter to the within-class scatter.[9].

The discrimination between the various classes is evaluated by the ratio of their projection between class scatter matrix (S_B) and the sum of their projected inside class scatter matrix (S_W) as stated by [10][11].

$$J(W) = \frac{|S_B|}{|S_W|} = \frac{|W^T S_B W|}{|W^T S_W W|} = [w_1, \dots, w_k] \quad \dots\dots\dots(2)$$

Where $[w_1, \dots, w_k]$ is the eigenvectors

$$S_B = \sum_{j=1}^c n_j (m_j - m)(m_j - m)^T, \quad \dots \dots \dots (3)$$

$$S_W = \sum_{j=1}^c \sum_{i=1}^{n_j} (x_{ji} - m_j)(x_{ji} - m_j)^T, \quad \dots \dots \dots (4)$$

Where c is the number of classes is number of samples in the class, x_{ji} is the i th sample in the class j , m is the mean of all classes, m_j is the mean of class j .

Such that: S_B : between – class scatter matrix, S_W : within – class scatter matrix,

W : projection matrix.

By increasing the Fisher criterion, we can obtain a special linear projection matrix, which increases within class changes and between class changes of the projected data. The final metric for separating two faces A and B is the Euclidean distance in the M dimension LDA space computed as [12].

$$D = \sqrt{\sum_{i=1}^M (y_i^A - y_i^B)^2} \quad \dots \dots \dots (5)$$

And by calculating the $(c - 1)$ Eigen vectors (w) and Eigen values (λ) of $(S^{-1}W S B)$.

The important concept of LDA method is to separate the class means of the projected directions well, at the same time determining a small change around these means as PCA method , the extracted characters of LDA are linear combinations of the original data. As LDA decreases the data efficiently on a low dimensional space, it is suited for graphical representation of the data sets [13].

2.3 Feature Selection Cuckoo Search (CS)

For simplicity in describing our new Cuckoo Search [14,15], we now use the following three idealized rules:

- Each cuckoo lays one egg at a time, and dumps it in a randomly chosen nest;
- The best nests with high quality of eggs (solutions) will carry over to the next generations.
- The number of available host nests is fixed, and a host can discover an alien egg with a probability $p_a \in [0, 1]$. In this case, the host bird can either throw the egg away or abandon the nest so as to build a completely new nest in a new location.

For simplicity, this last assumption can be approximated by a fraction α of the n nests being replaced by new nests (with new random solutions at new locations).

For a maximization problem, the quality or fitness of a solution can simply be proportional to the objective function. Other forms of fitness can be defined in a similar way to the fitness function in genetic algorithms[16].

When generating new solutions $x(t+1)$ for, say cuckoo i , a L'evy flight is performed

$$x^{(t+1)}_i = x^{(t)}_i + \alpha \oplus L'evy(\lambda), \dots \quad (6)$$

where $\lambda > 0$ is the step size which should be related to the scales of the problem of interest. In most cases, we can use $\alpha = O(1)$. The product \oplus means entry-wise multiplications. L'evy flights essentially provide a random walk while their random steps are drawn from a L'evy distribution for large steps

$$L'evy _ u = t - \lambda, \quad (1 < \lambda \leq 3), \dots \quad (7)$$

which has an infinite variance with an infinite mean. Here the consecutive jumps/steps of a cuckoo essentially form a random walk process which obeys a power-law step-length distribution with a heavy tail.

2.4 Face classification by Using Supporting Vector Machines

The sub-space approach in which the linear SVM is arranged is presented as a filter for the production of a sub-space, which is then used by the non-linear SVM compiler with the RBF nucleus to reveal the face [17].

SVM is a new way to classify both linear and nonlinear data. SVM algorithm can be described as follows: Nonlinear mapping is used to convert the original training data to a higher dimension. In this new dimension limits to the decision to separate twins from one category to another. SVM finds the data separated from two categories. This over plane is overloaded with support carriers, "drill" and "support vector" margins. The SVMs can do either prediction or classification [18].

The simplest case of a two-class problem where the layers are detachable linearly. Let the dataset D given as $(x_1, y_1), (x_2, y_2), \dots, (X | D |, y | D |)$, where x_i is a set of exercises with associated class descriptions, y_i . Each y_i can take one of two values, either +1 which corresponds to the categories by the computer buys = Yes or -1, which corresponds to the computer buys = not, respectively. Let's look at an example based on two inputs, A_1 and A_2 , It can be seen from the figure that 2-D data is linearly separated because a straight line can be plotted to separate each class +1 from each of the class-1 tuples. An infinite number of separation lines can be drawn; the best one is

that the target can be found, that is, one that will have a minimum error rating on an unprecedented seasoning. There is a technical problem with SVM technique for a maximum over plane margin [19].

However, hyper plane with greater margin is expected to be more accurate to classify the seasoning of future data from hyper plane with a smaller margin. This is why SVM is looking for an excessive plane with a larger margin, that is, Maximum Marginal Hyper plane (MMH). The unofficial definition of the margin that the shortest distance from an excessive plane to one side of its margin equals the shortest distance from the excessive plane to the other side of its margin, the hyper plane can be separated as follows:

$$W \cdot X + b = 0 \quad \dots \dots \dots \quad (8)$$

Where W is the weight vector, which is $W = \{w_1, w_2, \dots, w_n\}$; n is the number of attributes; and b is the numerical, often referred to as bias. In Figure (2.8), it shows only two possible superclass separation and associated margins. The best one is the one with larger margin should have greater circular accuracy. The sides of the margin can be written as in [20].

3. The Proposed System

The proposed system consist of training and testing phases as illustrated in Figure (2). In training phase several algorithms have been used to create dataset which will be use in the testing phase to decide right faces ,the training phase is based on the following stages :

- In Preprocessing stage used many methods to enhance the input images through applied convert color image to grayscale and Histogram Equalization.
- In Detection faces stage used Viola and Joins to detect multi-face in each input image .
- In Feature extraction stage, features vector in this stage well be extracted based on linear discernment analysis(LDA).
- In feature selection stage , to select best feature from feature extracted applied cuckoo search.

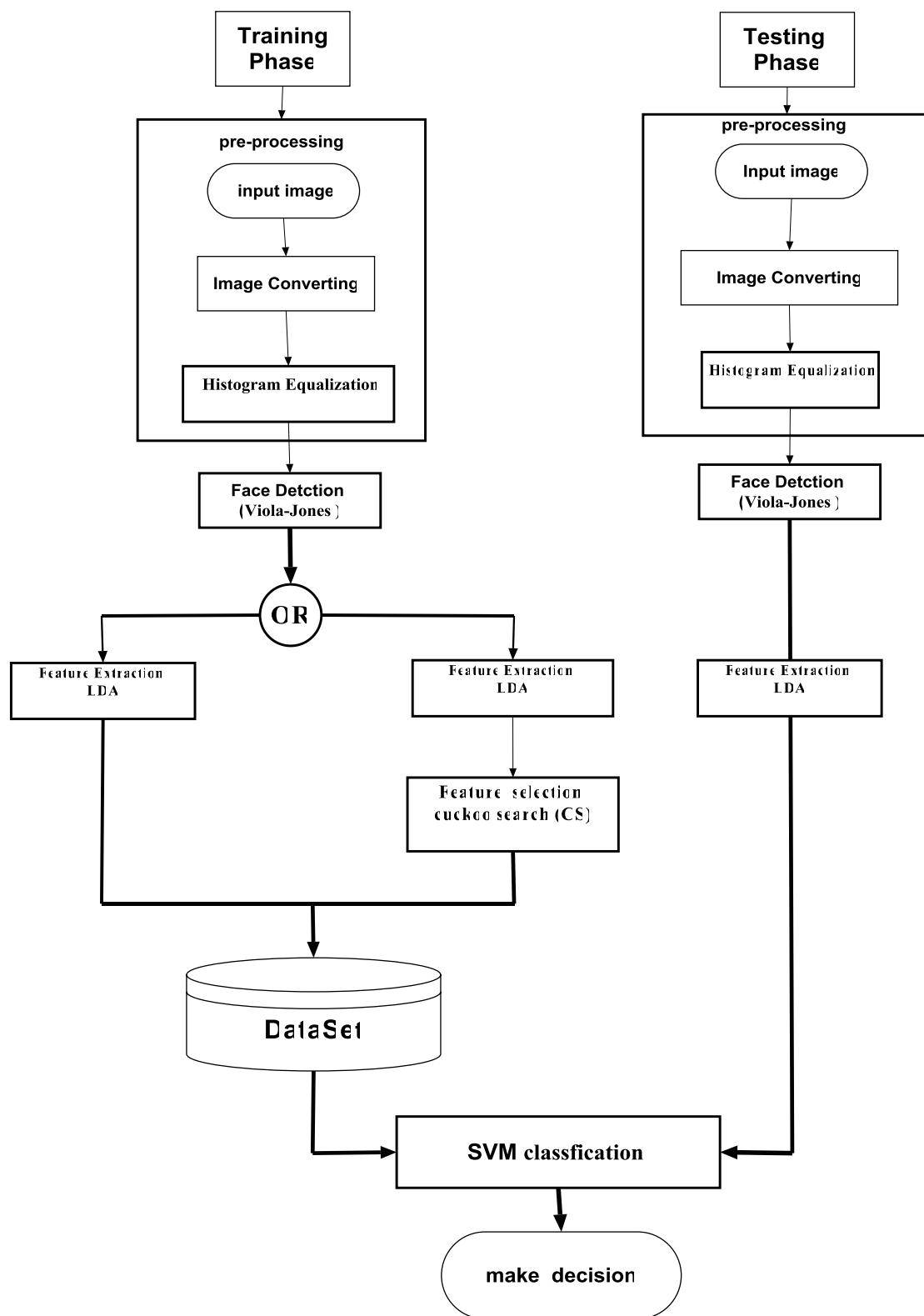


Figure (2): Proposed Face Recognition System

- In recognition stage have been support vector machine (SVM) to recognized the face or non-face.

In testing phase, all stage which applied in training phase is used this phase excepted feature selection stage.

4. Face Image Data Set.

MUCT database (figure.3) (MUCT stands for "Marlboro University of Cape Town") are used to consider system performance. In the MUCT database, 240 images of each person 12 pose variation image consist of lighting ,dark, smile, anger, skin black , and rotating , used in training phase 8 image each of person and testing phase 4image of person. The system display that increasing the many of training images can increase the recognition rate. The Viola- Jones method is used to detect the face on each database. This method has improved a high detection rate and all images have been detected and cut into databases. After being classified as "unknown," facial images can be added to a library (or to a database) with their element vectors for subsequent comparisons.

4.1 Training stage

The result of training phase include four dataset which content feature vector four face , nose , mouth and eye segments .each dataset was used separately in the testing phase to demonstrate the possibility of testing each segment of the face has been extracted.

This phase consist of the stages : read image , image converting, histogram equalization , face detection , feature extraction and feature selection ,each stage include many steps was it will explain in the following :

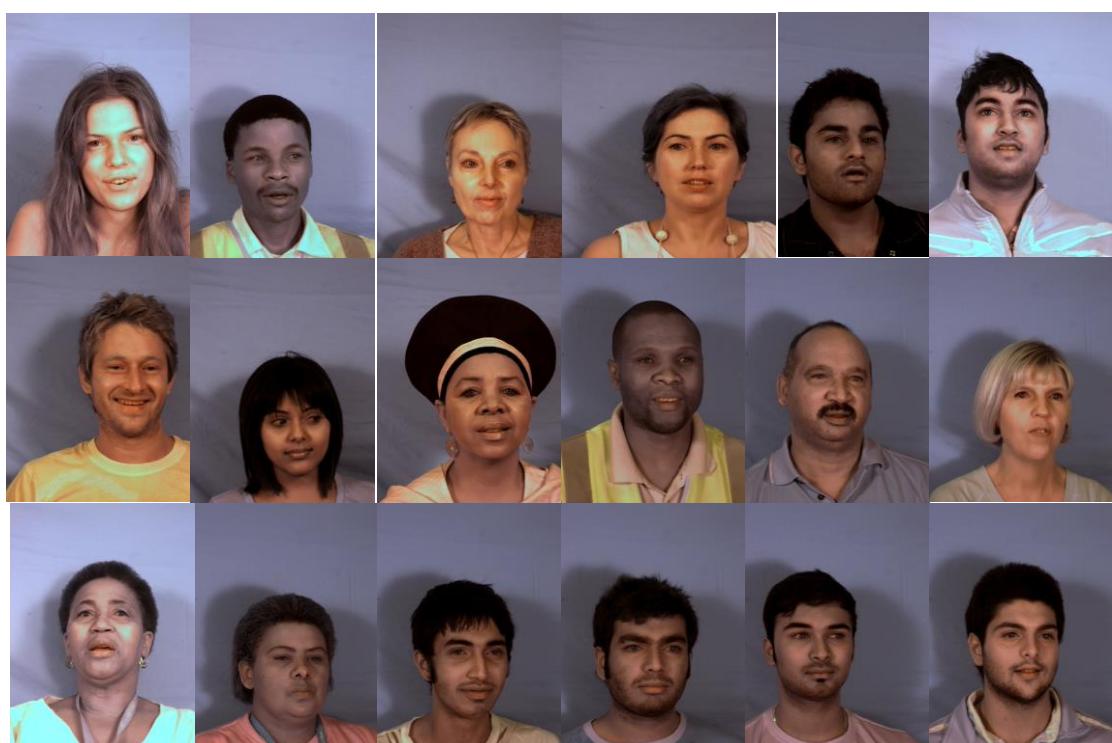


Figure (3): (MUCT) Database Face Images

4.2 Read image

The RGB color image is read as JPG image with resolution(70wight * 70high), these image taken 24 bit\ pixel ,the image data is separated into three band are Red ,Green and Blue .

4.3 Image Converting

The color image in converted to grayscale by using question:

$$\text{Grayscale value} = 0.2125R + 0.0715G + 0.722B \quad \dots\dots\dots (8)$$

4.4 Histogram equlization

Histogram Equalization is usually performed on low contrast images to improve image quality and face recognition performance. It changes the dynamic range (contrast range) of the image then so a result, some important facial features become more apparent[22].

The Histogram Equalization can be expressed mathematically as follows:

$$S_k = T(r_k) = \sum_{j=0}^k n_j / n \quad \dots\dots\dots (9)$$

Whereas $k=0, 1, 2 \dots L-1$

Here in Histogram Equalization (3.3) 'n' is the total number of pixels in an image, ' n_j ' is the number of pixels with gray ' r_k ' level, and 'L' is the total number of gray levels in the face image. The end result afterwards applying histogram equalization according to a pattern rear image is shown of Figure (4) Histogram Equalization. The Histogram Equalization on the left is from the original face image (between 6-250) and one on the right is after applying the Histogram Equalization. Figure (3.4) Image graph before and after Histogram Equalization.

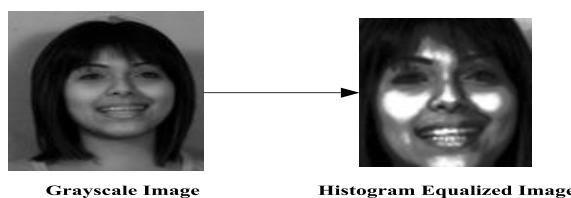
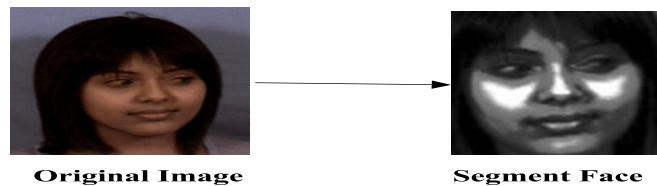


Figure (5): Histogram Equalization

First Phase: segment dataset into four segment consist of face, eye , noise and mouth

- Face segment image



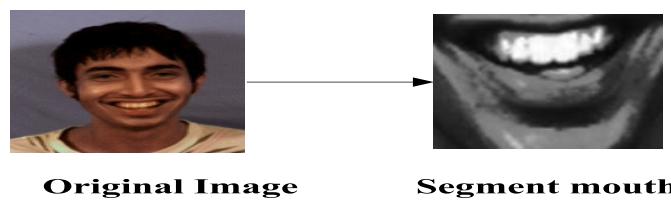
Figure(6): Face Segment Image

- Eye segment image



Figure(7): Eye Segment Image

- Mouth segment image



Figure(8): Mouth Segment Image

- Noise segment image



Figure (9): Noise Segment Image

5. Feature extraction

In this proposed system was extracted a set of features of the eye, mouth, face and noise regions which have been extracted by using algorithm LDA which illustrated in algorithm. This algorithm using testing and training samples
Linear discrimination analysis (LDA) reveals vectors in the basic space that best distinguishes between layers. For all samples of all strata, the scatter matrix S_T is determined between layer and the scatter matrix S_w within the layer.

6. Feature selection by Cuckoo Search (CS) Algorithm Optimization

Feature selection (FS) is a global optimization problem in machine learning, which reduces the number of features, removes irrelevant, noisy and redundant data, and results in acceptable recognition accuracy. It is the most important step that affects the performance of a pattern recognition system. There are many algorithms in face recognition, in this thesis getting 19 features of each image of each person. Total features are 3040. In feature selection, reduce the feature to better features of each image of the person into total numbers of features are 380. The best feature is applied in cuckoo search as following algorithm:

Algorithm1: Cuckoo Search(CS)

```
Objective function f(x), x = (x1, ..., xd) T ;  
Initial a population of n host nests xi (i = 1, 2, ..., n);  
while (t < MaxGeneration) or (stop criterion);  
    Get a cuckoo (say i) randomly by L'evy flights;  
    Evaluate its quality/fitness Fi ;  
    Choose a nest among n (say j) randomly;  
    If (Fi > Fj),  
        Replace j by the new solution;  
    End  
    Abandon a fraction (pa) of worse nests  
        [and build new ones at new locations via L'evy flights];  
    Keep the best solutions (or nests with quality solutions);  
    Rank the solutions and find the current best;  
end while  
Postprocessor results and visualization ;
```

7. Face Classification Method by Using SVM.

The training group of face images is contained attribute value represented by Eigenvector image class for a large number of people. To learn these attributes are considered as input to face classifier model. Testing the new person depending on face classifier mode.

In this work, the facial classifier model is used by SVM based on training faces images after applying feature extraction by using LDA then using feature selection by cuckoo search to select the best feature of each image. In testing phase, the face classifier model is used to inspect a new face image which does not belong to the training set. Classifier

face model gives a report (decision), SVM is a new method for the classification of both linear and nonlinear data. It uses a nonlinear mapping to transform the original training-data into a higher dimension. In this new dimension a for a decision boundary for separation of the tuples of one class from another. The SVM finds a data that was separated from two classes, this hyperplane using support vectors “training tuples” and margins “support vectors”. The SVMs can do either a prediction or a classification.

SVM is used in this work since the Training group data has two classes. SVM classifies face samples through determining the best hyper plane that withdraws for all data points of same class from those of the other class. The better hyper plane for SVM is selected based on the biggest margin between the two classes. Set of images are examined from different dataset to inspect of image for SVM classifier.

8. RESULT

The proposed system implemented on OPEN-CV C++ and JAVA language software and under Microsoft Windows environment. The databases (MUCT) is used to evaluate the system performance. In MUCT database, 240 image of four segment to training and testing images for each segment is used. Viola-Jones method is used for face detection on each database In this thesis the method of viola and jones was adopted because it is characterized by speed processing and high accuracy by applying more than one algorithm In this thesis, this method was used for 160 images as training samples to detect four segment of the face in addition to detection more than one face in the same image as Detect of face region ,Detect of eye region ,Detect of mouth region ,Detect of noise region and Detect of multi-face in the same image .

However, LDA is applied on the detected cropped images for feature extraction and dimension reduction. Different number of training and testing images are used in each database In this proposed system was extracted a set of features of the eye , mouth , face and noise regions which have been extracted by using algorithm LDA. this algorithm using testing and training samples Linear discrimination analysis (LDA) reveals vectors in the basic space that best distinguishes between layers. For all samples of all strata, the scatter matrix S_T is determined between layer and the scatter matrix S_w within the layer.

Then applied on feature selection cuckoo search algorithm(CS) to select better feature and to improved faster algorithm and reduce size Feature selection (FS) is a global optimization problem in machine learning, which reduces the number of features, removes irrelevant, noisy and redundant data, and results in acceptable recognition

accuracy. It is the most important step that affects the performance of a pattern recognition system. there are many algorithm in face recognition , in thesis getting on 19 features of each image of each person total of features are 3040 in feature selection reduce the feature to better feature of each image of the person into total numbers of feature are 380 the best feature this applied in cuckoo search after done used Support vector machine (SVM) is used to classification training group of face images is contained attribute value represented by Eigenvector image class for a large number of people. To learn these attributes are considered as input to face classifier model. Testing the new person depending on face classifier mode.

Accuracy

$$= \frac{\text{Total Number of sample} - l \text{ Number of sample that False accepte}}{\text{Total Number of Sample}} * 100\%$$

LDA	Datase t	Training set	Testing set	hit	miss	accuracy
face	240	160	77	76	1	99%
nose	240	160	75	71	4	95%
mouth	240	160	80	77	3	96%
Eye	240	151	66	62	4	94%

Table1: Recognition results of the proposed LDA Algorithm

The result Recognition results of the proposed LDA Algorithm in table1: data set 240 samples to doing 160 training phase and 80 testing phase ,in training phase after operation detection face by using voila-jones of numbers of faces and segmentation into four regions eye , face , mouth and noise , after this phase using feature extraction b using algorithm LDA getting to 3040 feature of image this mean each person have 8 samples from 12 sample of training has been extracted each image 19 feature $(8*20*19)=3040$ feature to face training face accuracy 99% , eye training face accuracy 94% , mouth training face accuracy 96% and noise training face accuracy 95% .

cuckoo	Datase t	Training set	Testing set	hit	miss	accuracy
face	240	160	20	77	0	100%
nose	240	160	20	75	0	100%
mouth	240	160	20	80	0	100%
Eye	240	151	20	66	2	97%

Table2: Recognition results of the proposed Cuckoo Algorithm

The recognition results of the proposed Cuckoo Algorithm in table2: data set 240 samples to doing 160 training phase and 80 testing phase ,in training phase after operation detection face by using voila-jones of numbers of faces and segmentation into four regions eye , face , mouth and noise , after this phase using feature extraction b using algorithm LDA getting to 3040 feature of image this mean each person have 8 samples from 12 sample of training has been extracted each image 19 feature $(8*20*19)=3040$ feature , in feature selection this 3040 feature selected better feature each image to getting 380 feature this mean $(19*20*1)=380$ feature , to face training face accuracy 100% ,eye training face accuracy 97% because 2 miss from image the system cannot detected , mouth training face accuracy 100% and noise training face accuracy 100 %.

9. Conclusions

The suggested algorithm is used to classify the face image whether it is (known or unknown). This algorithm has a vital role in surveillance and authentication systems. The main conclusions of the proposed algorithm is that the face image recognition is a full-face display of the digital image by applying the Voila-Jones algorithm that is used to detect face image and it is a fast method. In addition, to find an image of low dimensional, redundant and noise information are determined by the use of LDA for obtaining the value of the feature. The algorithms which are used for feature selection methods in this research are reduced the numbers of features and increase the rate of recognition. In conclusion, the obtained results indicate that SVM without cuckoo search is provided 98% of accuracy while SVM with cuckoo search gives 100% of accuracy for the same sample of image data sets.

REFERENCES

1. Rouhi, Rahimeh, et al. "A review on feature extraction techniques in face recognition." *Signal & Image Processing* 3.6 (2012): 1.
2. Schalkoff, Robert J. *Digital image processing and computer vision*. Vol. 286. New York: Wiley, 1989.
3. Nazir, M., et al. "Feature selection for efficient gender classification." *Proceedings of the 11th WSEAS International Conference*. 2010.
4. Viola Paul, and Michael Jones. "Robust real-time face detection." *International journal of computer vision* 57.2 (2004): 137-154.
5. Barnouti, Nawaf Hazim, et al. "Face Detection and Recognition Using Viola-Jones with PCA-LDA and Square Euclidean Distance." *International Journal of Advanced Computer Science and Applications (IJACSA)* 7.5 (2016).
6. Li, Stan Z., and K. Anil. "Jain. *Handbook of Face Recognition*." (2005).
7. Nain, Neeta, et al. "Face recognition using pca and lda with singular value decomposition (svd) using 2dlda." *Proceedings of the World Congress on Engineering*. Vol. 1. 2008.
8. Ye, Fei, Zhiping Shi, and Zhongzhi Shi. "A comparative study of PCA, LDA and Kernel LDA for image classification." *Ubiquitous Virtual Reality*, 2009. ISUVR'09. International Symposium on. IEEE, 2009.
9. Kakadiaris, Ioannis A., et al. "Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.4 (2007).
10. Marcialis, Gian Luca, and Fabio Roli. "Fusion of LDA and PCA for Face Verification." *Biometric Authentication* 2359 (2002): 30-37.
11. Martínez, Aleix M., and Avinash C. Kak. "Pca versus lda." *IEEE transactions on pattern analysis and machine intelligence* 23.2 (2001): 228-233.
12. Yu, Hua, and Jie Yang. "A direct LDA algorithm for high-dimensional data—with application to face recognition." *Pattern recognition* 34.10 (2001).
13. Liu, Chengjun, and Harry Wechsler. "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition." *IEEE Transactions on Image processing* 11.4 (2002): 467-476.
14. Yang, Xin-She, and Suash Deb. "Cuckoo search via Lévy flights." *Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on*. IEEE, 2009.
15. Yildiz, Ali R. "Cuckoo search algorithm for the selection of optimal machining parameters in milling operations." *The International Journal of Advanced*

- Manufacturing Technology (2013): 1-7.
16. Gandomi, Amir Hossein, et al. "Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems." Engineering with computers 29.1 (2013).
 17. Shen Maodong and Cao Jiangtao,Li Ping,"Independent Component Analysis for Face Recognition Based on Two Dimension Symmetrical Image Matrix", IEEE, 978-1-4577-2073-4,2012
 18. Joachims, Thorsten. "Text categorization with support vector machines: Learning with many relevant features." Machine learning: ECML-98 (1998).
 19. Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.
 20. Franc, Vojtěch, and Soeren Sonnenburg. "Optimized cutting plane algorithm for support vector machines." Proceedings of the 25th international conference on Machine learning. ACM, 2008.

Generic Synthesis System Based on Agile Methodology for Multimedia Mobile Web Learning Modules

Linda lafta Gashim
Department of Computer Science
College of Science / Mustansiriyah University
Baghdad, Iraq
lindalafta85@gmail.com

Assist. Prof. Dr. Karim Qasim Hussein
Department of Computer Science
College of Science / Mustansiriyah University
Baghdad, Iraq
karimzzm@yahoo.com

Abstract—The main idea is to study invest advantage of multimedia elements to producing educational in web mobile platform in efficiency to meet the specific need for student and instructor at less complex for all device. Web Mobile application learning has been recognized a new approach in information teaching, emerged as a new and promising learning modality and providing more interactivity and flexibility to learners, student and instructor in carrying out educational activities and practices. The proposed model is also enriched based on agile (scrum framework) for manage system development. Also, it is iterative and incremental approach, this model consists number of sprints; sprints are series of development phases that finished the subsystem incrementally. The design of the proposed system is performed using object oriented concepts and the implementation is done using Asp.net studio and C# language with SQLServer database to store all required data for the system. Also, using The Planning Poker technique estimated the effort in terms of units of work referred to as ‘story points’ reflecting the complexity. The quality evaluation of Web Mobile learning environments, has been evaluated based on a set of ISO/IEC standards. The result of this present work is design system with two sprints, the sprint1 and Sprint2 is design and development Web Mobile for lesson and online exam.

I. Introduction

Learning become one of the closely essential activities in the existing knowledge, which is characterized by information age, globalization, knowledge acquisition and transfer, the information and communication technology revolution. [1]. This work has explored the concept and characteristics of web mobile learning, mobile devices and other related issues, which proposed design principles for web mobile learning system. The biggest advantage of using synthesis system for packages to create multimedia educational content is that interactivity can be incorporated into the content. In educational model, has been efficient for several reasons, it associated with constraints that could simplified the learning process, one of this constrain is the time where leaner can take the material at any time, anther constraint is space where leaner can access to material at any space in world [2]. In [3] presented framework for development web methodology based on agile method. Through this work show that this several Web application systems require to change some parts from the analyze beginning due to various reasons. This work suggest that Iterative and incremental development is the best method to settle this common trouble. And through this work Applying ISO - IEC9126 quality model to the framework and improve the efficiency and insure that the quality of Web development [3]. And in [4] introduced a system that design and development of a user interface for instructors to create exam and an interface that students

can access via a mobile platform and access the quiz, in this thesis apply one of the agile framework is Rapid Application Development(RAD).

II. Agile Methodology

Classically classification of software development frameworks there have been three forms of methodological frameworks: linear, iterative, and combination of both. The most common linear framework is Waterfall where projects consist of sequential phases with acceptance of some overlap. Every step in a waterfall process must be finished before moving on to the next. Iterative development is quite opposite, it excludes initial planning but focuses on constant changes, and stimulates continuous revision and improvement of software. The work is broken up into small pieces that are developed over some period and finally put together when they are ready, an example of iterative framework is Prototyping. Iterative frameworks can be also used in combination with linear methods, setting up such frameworks as Incremental, Spiral, Rapid application development (RAD), and Extreme Programming. Incremental approach improves development process; iterative approach increases product's quality Agile suitable software process model for Mobile Application development [7].

Agile is emerging Software engineering methodology based on feedback and embracing their changing needs. It is considering flexible approach where system phases are guided by product advantage. There are several agile frameworks like lean development, extreme programming, scrum, and dynamic system development methods which come under the agile methodology [8].

A. Scrum

Scrum is a framework for agile methodology software development that is iterative and incremental. Scrum can concentrate on describe how the team member should function in order the system flexibly. the mainly task of Scrum is to be used for management of software development process, it can be applied to run software maintenance teams, or as a general project management approach [6].

B. Process of Scrum

A Scrum development process as the one in Fig (1) illustrated the general structure of scrum consists of a number of sprints;

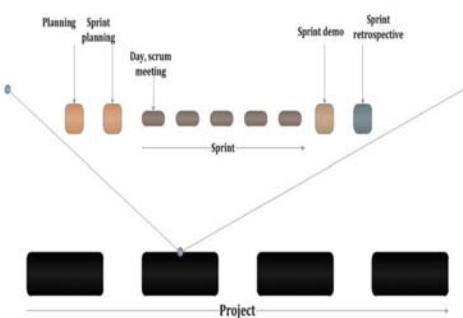


Fig (1): General Structure of the scrum process.

sprints are series of short development phases that delivers the product incrementally [9]. Fig (2) illustrate structure of the scrum process:

1. **Pre -phase: - includes two subs (planning, architecture / high level design).**
2. **Development**
3. **Post –development**

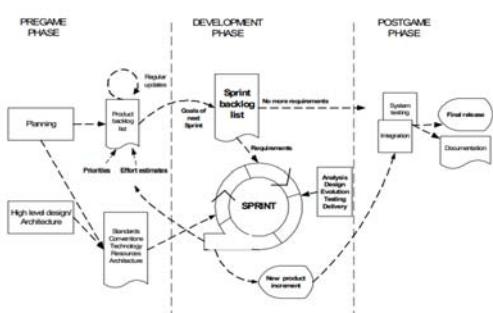


Fig (2): Scrum Process

1. **Pre -phase:** - pre-phase can divided in two sub phases as shown following:

Through Planning involved the definition of the software being developed. A product backlog list is created containing all the requirement that are currently known. the requirement could be produced from software developers. In this phase, the effort necessary for their achievement system is estimated and requirements prioritized for task is determined. the **product backlog** table is constantly updated with new and more detailed items, as well as with more accurate estimations and new priority orders. Planning also contain the definition of the project member, tool and other resource, risk assessment approval. through each iteration, the change in PB is reviewed by the scrum member so as to acquisition their commitment for the new iteration [10].

The product backlog is like a prioritized queue of tasks, its contains a list of the requirement, features and enhancements. PB include a characterization, a priority, and an estimation of the tasks. A Product Backlog (PB) contains customer requirements (including functional and non-functional), as well as technical requirements [12].

In scrum, it used user story. User story reflect quantity that expresses the amount of functionality and the complexity of the tasks. Every member creates their own story point sizing framework based on the type of work they do, the skills and experience of the team member, and what they personally perceive to be a small, medium, or large amount of work [13].

This paradigm of the user story is described in this technique: [As an end user role), I want (the desire) so that (the rationale).

The second level from pre-phase consider high level design of the project include the architecture is planned based on the tasks in the PB. in case of an enhancement to an existing system, the changes needed for achievement the Backlog items are determine along with the problems they may cause.

2. **The development phase:** - this phase is treated as a "black box" where the unpredictable is expected. the various environmental and technical variables (such as resources, implementation technologies, timeframe, quality, requirements and tools, and even development methods) identified in Scrum, which may change. In this phase, the system is developed in sprint, the sprint with within cycle is shown in Fig (3).

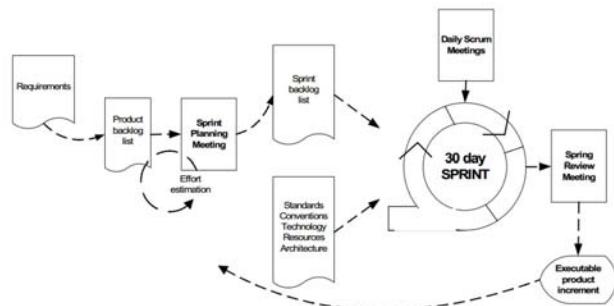


Fig (3): Sprint cycle

Sprints are frequently cycles where the tasks are developed or enhanced to produce new iteration. each Sprint includes the classical phases of software development process: requirements, analysis, design, evolution and delivery phases . Usually Sprint is determined to last from one week to one month. there may be, for example, four to eight Sprints in one project development process before the system is prepared for delivery. In addition, there may be more than one developer building the increment. A sprint is repeated cyclically until the system is finished. as shown in Fig (4).

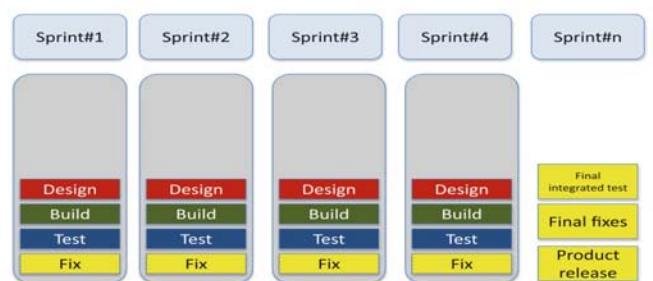


Fig (4): Sprint leading to final product release.

3. The post phase: - involve the closure of the release. Post phase is entered when an agreement has been made that the environmental variables such as the task are completed. The system is ready preparation for the release. Through this phase, implementation the tasks such as the integration, system testing and evaluation [10].

III. Mobile web Application Development

In mobile web system life cycle are developed using a methodology to make application implementation on many devices with various screen sizes. Mobile web architecture design based on different techniques like CSS, HTML and java script with used C# or visual basic. Through This type of applications attempted to combine the best of both approaches; by utilities the advantage of server computing but don't treat the apparatus only as a front end [14]. One of the most the JavaScript library bootstrap which is designed for modern browsers and smart devices. Bootstrap is the most popular HTML, CSS, and JS framework for developing responsive, mobile first projects on the web. It is an open source library of UI components developed by Twitter. The components are built using the responsive web design principles, which makes this library extremely valuable for web applications that needs to automatically adjust its layout depending on the screen resolution. This library exploited the advancement in HTML5, CSS3 and JavaScript and provide an Application Programming Interface (API) for developer to create web mobile-friendly applications. [15].

IV. Design and Implementation proposed System

In the proposed system is divided into two sprints, through the first sprint provided generic synthesis for lessons (text, video, image), secondly supply generic synthesis for online exam.

A. Planning and Estimation system

This section describes first phase and the initial planning for the system, the product owner in Planning phase determines the user stories for each iteration that fit the estimate of effort as established by the team in scrum. it is important to notice the estimation component and the high level of uncertainty at this point of the project. Table 1 Initial Plan for system represents the product backlog contain list of features or "user stories". Development team divided user stories into tasks through the phase of planning. The main tasks initially planned contain the initial and final date and the estimated effort. The time and effort are calculated using a 9 hours-a-day working calendar for each participant in the project. And then describe the estimation for each story in backlog implementation and desired functionality. when survey the product backlog they are focus on two things mainly. First, make sure that all the require specification are represented in backlogs. Second, verify that all the estimation for developing backlog is as precision as possible. Table (1): product backlog for proposed system.

B. Architecture - High Level Design

The backlog through the phase obviously defined is being made. In the next backlog, all the changes are identified in a new iteration. A repeated architecture is start generated to provided new contexts and additional requirements. For the system product backlog was created in Microsoft Excel as in in Table (1). This product backlog contains key activities for the project and will continually be updated once more stories are being produced. From the product backlog, the top stories will then be picked for development in the next sprint. The proposed system has two sprints as shown in Fig (5).

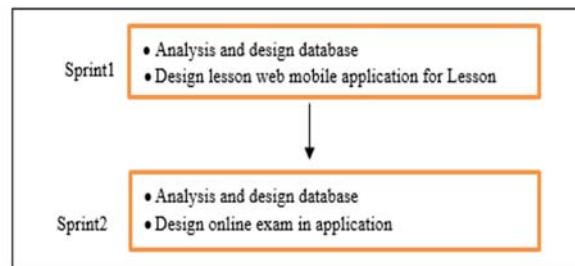


Fig (5): Sprints for system Covering Development from Initiation to a Finished.

C. Sprint 1 and Sprint 2

The general architecture of the system and how work within the whole environment will illustrated in the following Fig (6). This Fig explains that electronic learning services is server based architecture, there are two users in the system, as illustrated in the following: -

- 1-Instructor tier: runs in any computer or mobile platform that is responsible for upload data to the database.
- 2- Students tier: In the student side, the user of a web mobile application can view data across the internet and into the mobile application or computer platform.

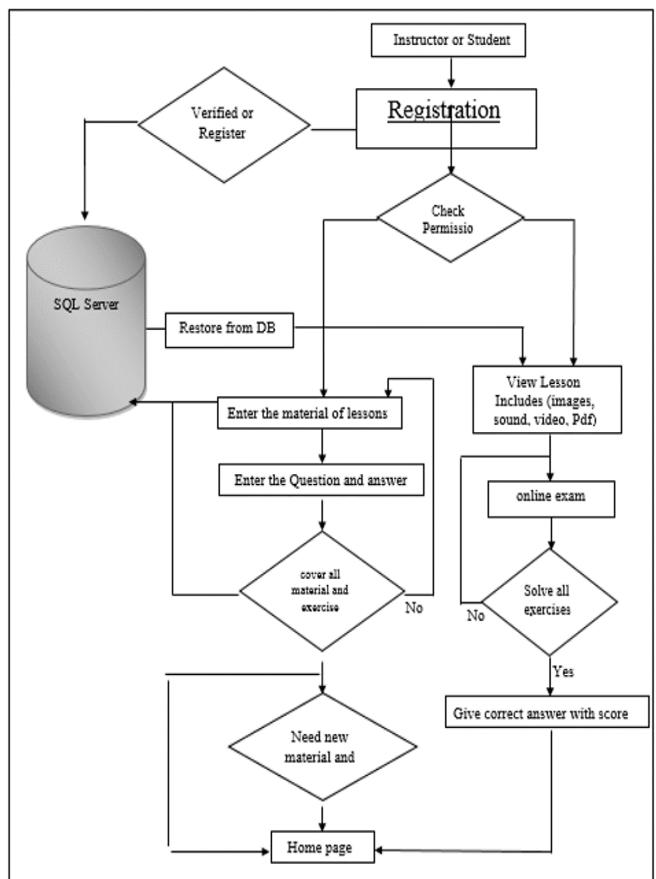


Fig (6): Architecture of proposed system

Table (1): product backlog for proposed system.

NO	User story	iteration	Task	complexity	Estimate on day)	No. of work (hour)	Effort (h)	Start Date	End Date
1	as an instructor, I want to register to system so I can use system	1	1. Design database in SQL server	Large	5	9	45	1/8/2016	5/8/2016
2		1	2. design tables for USER information.	Large	2	9	18	6/8/2016	7/8/2016
3		1	3 built Sign up and sign in interface.	Medium	3	9	27	8/8/2016	10/8/2016
4	as an instructor, I want to upload lesson so that student can view lesson.	1	design table for lesson details	Large	3	9	27	11/8/2016	13/8/2016
5		1	built interface in website to submit materials in website	Medium	3	9	27	14/8/2016	16/8/2016
6	as an instructor, I want to upload multimedia elements so that student can more active	1	design table for multimedia elements	Medium	2	9	18	17/8/2016	18/8/2016
7		1	built interface to upload multimedia elements and other resources	Large	4	9	36	19/8/2016	22/8/2016
8	as an instructor, I want to search button in system so that I can access to lesson	1	design interface for search	Small	2	9	18	23/8/2016	24/8/2016
9	as an instructor, I want to delete/update/view lesson so that I can manage system	1	design interface for delete, update and view lesson	Small	2	9	18	25/8/2016	26/8/2016
10	as an instructor, I want to create exams online so that student can take test	2	1-design tables for test, subject and question.	Large	5	9	45	3/9/2016	10/9/2016
11		2	2-build interfaces for test, subject and question in website	Large	5	9	45	11/9/2016	15/9/2016
12	as an instructor, I want view result of true answer of test in online exam so that can have evaluated student.	2	design tables for result and student.	Medium	3	9	27	16/9/2016	18/9/2016
13		2	design interface for true answer and result report of test in web site	Small	2	9	18	19/9/2016	20/9/2016
					41		369		

D. Analysis and Design of proposed system

In the Fig (5) show that the sprint1 and sprint2 achieve to build design database, built mobile web application, system consists of the following parts:

1. Web Forms: Web form can be divided into two categories.

First, Mobile Web Form: The documents that sends output to the user. secondly, **Web services:** The documents that do not send output to the user, only checks the information that are input from user.

2. Database Server: The web forms are implemented using the SQL server database.

analysis of learning content and learner by specifying instructional objective, knowledge and skill for collaborative learning. A Use Case is a way to understand and describe the requirements. The following section give a brief description of the main use cases of the MMLS system. Fig (7) shows the diagram of the general use cases.

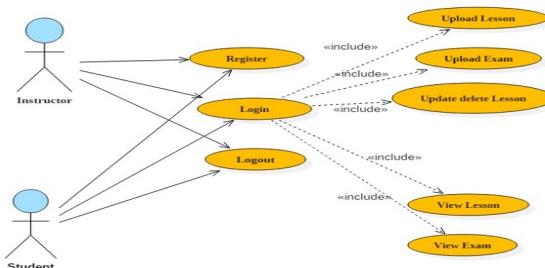


Fig (7): Use case Diagram for proposed system

V. System Implementation and Testing

Choosing the appropriate language for implementation of any system is very critical concept. In the implementation of system, SQL server with Asp.net and c#, html and java script language in mobile web was chosen for the greatest features they provide to develop complex systems perfectly; such as debugging tools, publishing tools

and other fascinating features, Fig (8)(9) show some interface for proposed system

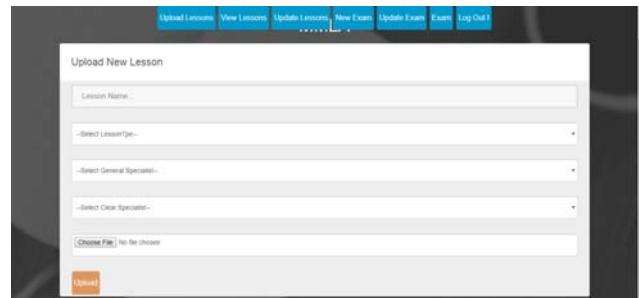


Fig (8) interface for upload lesson

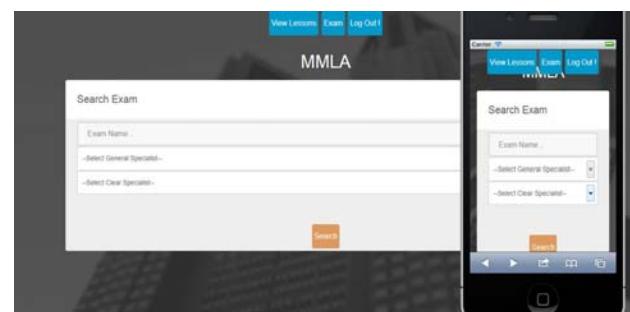


Fig (9) interface for Online exam

VI. Different Effort Estimation in Agile Methodology

The differerance between the estamtion time in the planning phase in agile development and real effort shown in the Table (2) for system with relative error . for each story of each sprint, the effort estimated with Story Points. The real effort, and the relative error of the Story Points estimates was calculated as follows [59]:-

$$\text{Relative error: RE} = \frac{\text{Actual} - \text{Estimate}}{\text{Actual}}$$

The relative error (RE) gives an indication of the divergence between the values estimated by the model and the actual values, expressed as a percentage. This relative error can be either positive or negative, representing either an overestimation or an underestimation. Fig (10) Estmation and actual hours for system(N=13 tasks) with story point.

Table (2): Estimation and actual hours for system(N=13 tasks)

NO	iteration	No of Work(hours)	Effort (estimation) (1)	Effort (actual) (2)	under/over estimation (H) (3) = (2)-(1)	Relative Error (4) = (2)-(1)\(2)
1	1	9	45	40	-5	-13%
2	1	9	18	14	-4	-29%
3	1	9	27	25	-2	-8%
4	1	9	27	25	-2	-8%
5	1	9	27	30	3	10%
6	1	9	18	15	-3	-20%
7	1	9	36	30	-6	-20%
8	1	9	18	15	-3	-20%
9	1	9	18	20	2	10%
10	2	9	45	43	-2	-5%
11	2	9	45	40	-5	-13%
12	2	9	27	25	-2	-8%
13	2	9	18	15	-3	-20%
Total		369	337	-32		

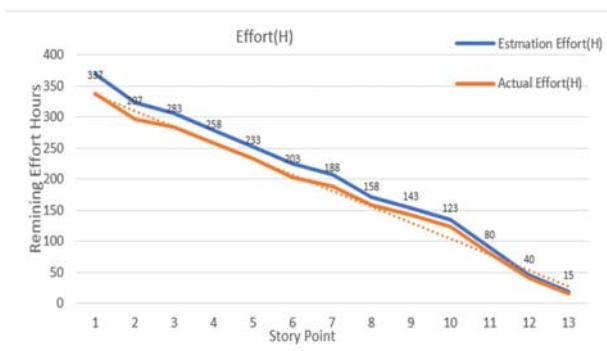


Fig (10): Estimation and actual hours for system.

VII. Quality Evaluation for System

The method of evaluation of system based on method based on a set of ISO - IEC standards [16]. A quality model consists of a group of characteristics that relate to each other and which supply the basis for specifying quality requirements and quality evaluation.

Table (3): Score for each Quilty Criteria for system. And Fig (11) illustrated Score for each Quilty Criteria in sytem.

Question	functionality	Security	performance	Pedagogical	usability	support	communication	portability
1	10	10	0	10	10	10	10	10
2	10	10	7.5	10	10	10	0	0
3	7.5	10	10	10	7.5	10	10	10
4	10	7.5	7.5	10	10	7.5	10	10
5	10	10	7.5	10	7.5	0	7.5	10
6	10	10	10	10	7.5	7.5	7.5	10
7	10	7.5	7.5	7.5	10	10	5	10
7	7.5	7.5	7.5	10	10	7.5	0	10
9	10	7.5	10	0	10	7.5	9	10
10	0	10	10	10	7.5	10	6	10
Score	85	90	77.5	87.5	90	80	65	90

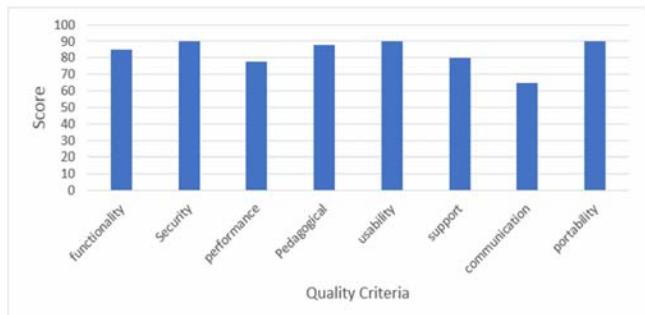


Fig (11): Overall Score for each Quilty Criteria in system.

VIII. Conclusion

System is a high-quality Web mobile development method, combined the advantage of Agile Development and reused Web framework. It is an excellent Web development technique for the Web mobile application system which require to quickly service, quick response, and rapidly adapt the change in requirement. Web mobile learning system is important in future learning system, can be easily adopted by institutions to be used as a separate learning system due to use of user friendly menus and easy to access functions and features. Client-Server network architecture provides more control and easy management for the user however at the expense of scalability and congestion problems. quality analysis system based on the quality factors of ISO-IEC9126 quality model. The results show that improve the efficiency of the system.

Acknowledgement

I would like to give special thanks to my supervisor. **Dr.Karim Qasim** for his valuable advice, guidance during the study research. My deep thanks also to Dr. Sajida Lafta, Dr. Lamia Lafta and Khawlaa Turkey in university of Baghdad for support during the period of study.

Reference

- [1] Firouz B. Anaraki "A Web Based Learning System for Learning English on Mobile Devices KMITL ", Journal of Sci. Tech. , Vol. 14, No. 2 , Jul. - Dec. 2014.
- [2] Ken Neo T. K. & Mai Neo, "Interactive multimedia education: Using Author ware as an instructional tool to enhance teaching and learning in the Malaysian classroom ", Journal of Interactive Educational Multimedia, No.5 , pp 80 - 94, October 2002.
- [3] Hu Ran, "Agile Web Development with Web Framework"; 2008 IEEE.
- [4] Andrew Montanaro "IMPLEMENTING MINI QUIZZES TO INCREASE STUDENT LEARNING FOR A CLASS VIA A MOBILE LEARNING APPLICATION " Department of Computer Science, University of North Carolina Wilmington, 2012.
- [5] Farrukh Shahzad," Modern and Responsive Mobile-enabled Web Applications", Procedia Computer Science 110 (2017) 410–415.

- [6] Vinicius Pereira and Antonio Francisco do Prado, "Introducing a New Agile Development for Web Applications Using a Groupware as Example" journal of E.R. Hruschka Junior et al. (Eds.): INTECH 2011, CCIS 165, pp. 144–160, 2011.
- [7] Tatjana Pavlenko "Applying Agile Methodologies to Design and Programming", Thesis, Tallinn University Institute of Informatics, 2012
- [8] Asra Khalid, Sobia Zahra, Muhammad Fahad Khan "Suitability and Contribution of Agile Methods in Mobile Software Development" International Journal of Modern Education and Computer Science, Vo. 2, PP. 56-62. 2014.
- [9] Emelie Jonsson "Agile Development and User-Centered Design-a case study at Sony Mobile Communications AB" Master thesis at the department of Design Science, LTH,2013.
- [10] Murali Chemuturi "Requirements Engineering and Management for Software Development Projects" Springer Science and Business Media New York 2013.
- [11] Marko Seikola "The Scrum Product Backlog as a Tool for Steering the Product Development in a Large-Scale Organization "Master's thesis for the degree of Master of Science in Technology submitted for inspection. Espoo, 7th May 2010.
- [12] Imrul Kayes1 · Mithun Sarker2 · Jacob Chakareski3" Product backlog rating: a case study on measuring test quality in scrum" Innovations software engineering PP. 303–317, (2016).
- [13] Bo Chen1,2, Hui He3, and Ying Zhang," A Hybrid Recommendation Model for HTML5 Mobile Web Applications" Springer, pp. 638–647, 2012.
- [14] Mario E. Moreira " Being Agile",book, Apress.
- [15] Mohamad Nizam AYUB, Santhimathy T. VENUGOPAL, Nurul Fazmidar Mohd NOR, "Development of Multimedia Authoring Tool for Educational Material Disseminations"; Journal of Informatics in Education, Vo. 4, No. 1, PP. 5–18,2005.
- [16] Gustavo Willians Soad, Nemesio F. Duarte Filho and Ellen Francine Barbosa "Quality Evaluation of Mobile Learning Applications" Frontiers in Education Conference (FIE) 2016 IEEE, pp. 1-8, 2016.

More than 30 papers have been published in the area of interest

AUTHORS PROFILE



Name: Linda Lafta Gashim

Nationality: Iraqi

Graduated a High Master degree in
Computer Science.

Main interest is software Engineering,
Multimedia, Web mobile Design.



Name: Assist. Prof. Dr. Karim Q. Hussein

Nationality: Iraqi

Ph.D. in Computer Science.

Department of Computer Science – Faculty of
Science – Mustansiriyah University Baghdad-
Iraq.

Main Area of research and teaching
(Multimedia ,3D animation, e-learning
,Website Technique and Mobile Computing)

Data Mining Methods for Worm Detection Using Variable Length Instruction Sequences

Muazzam Ahmed Siddiqui ¹, Safaa Alkatheri ²

Department of Information Systems, Faculty of Computing & Information Technology, King Abdulaziz University

¹maasiddiqui@kau.edu.sa

²ssalehalkatheri@stu.kau.edu.sa

Abstract—A Worm is a standalone malware that replicates itself and does not need a host to propagate. The prevailing detection approach uses fixed size sequence of characters extracted from the worm as a signature. Although very fast, the signature detection approach is ineffective for zero day attacks. Building upon our previous work to use data mining techniques for malware detection, we present a comparison of variable length instruction sequence to fixed size n-grams to detect worms. We define a sequence as a series of instructions till a branch is encountered. Our results indicated that using the variable length sequences as features to a machine learning classifier provided better results than using fixed length n-grams. We built and compared several tree based classifiers and were able to achieve 95.6% detection rate on novel worms using the variable length instruction sequences.

Index Terms—Data Mining, Worm Detection, Binary Classification, Static Analysis, Disassembly, Instruction Sequences

I. INTRODUCTION

Computer virus detection has evolved into malware detection since Cohen first formalized the term computer virus in 1983 [1]. Malicious programs, commonly termed as malwares, can be classified into virus, worms, trojans, spywares, adwares and a variety of other classes and subclasses that sometimes overlap and blur the boundaries among these groups [2]. The most common detection method is the signature based detection that makes the core of every commercial anti-virus program. To avoid detection by the traditional signature based algorithms, a number of stealth techniques have been developed by the malware writers. The inability of traditional signature based detection approaches to catch these new breed of malwares has shifted the focus of malware research to find more generalized and scalable features that can identify malicious behavior as a process instead of a single static signature.

The analysis can roughly be divided into static and dynamic analysis. In the static analysis the code of the program is examined without actually running the program while in dynamic analysis the program is executed in a real or virtual environment. The static analysis, while free from the execution overhead, has its limitation when there is a dynamic decision point in the programs control flow. Dynamic analysis monitors the execution of program to identify behavior that might be deemed malicious. These two approaches are combined also [3] where dynamic analysis is applied only at the decision-making points in the program control flow.

In this paper we present a static analysis method using data mining techniques to automatically extract behavior from worms and clean programs. We introduce the idea of using sequence of instructions extracted from the disassembly of worms and clean programs as the primary classification feature. Unlike fixed length instructions or n-grams, the variable length instructions inherently capture the programs control flow information as each sequence reflects a control flow block.

The difference among our approach and other static analysis approaches mentioned in the related research section are as follows.

First, the proposed approach applied data mining as a complete process from data preparation to model building. Although data preparation is a very important step in a data mining process, almost all existing static analysis techniques mentioned in the related research section did not discuss this step in detail except [4]. Second, all features were sequences of instructions extracted by the disassembly instead of using fixed length of bytes such as n-gram. The advantages are:

- 1) The instruction sequences include program control flow information, not present in n-grams.
- 2) The instruction sequences capture information from the program at a semantic level rather than syntactic level.
- 3) These instruction sequences can be traced back to their original location in the program for further analysis of their associated operations.
- 4) A significant number of sequences that appeared in only clean program or worms can be eliminated to speed up the modeling process.
- 5) The classifier obtained can achieve 95% detection rate for new and unseen worms.
- 6) Instruction sequences are a domain-independent feature and the technique can be used without any modification to detect other malwares e.g. virus, trojans, spywares etc.

It is worth noting that a dataset prepared for a neural network classifier might not be suitable for other data mining techniques such as decision tree or random forest.

II. RELATED RESEARCH

Malware is a great threat to users and systems. They spread across the internet and causing damage to computers. One of the popular techniques to detect malware is signature

detection. This technique matches the executables against a unique signature to detect malware. However, some of detecting malware techniques are not effective against the previously unknown pattern and complicated malware. Due to that reason, the researchers are developing more advanced methods to detect them by using data mining techniques instead of depending on the signature methods.

This section discusses some related works of malware detection and clustering by using data mining techniques. We categorized the studies of malware detection by using data mining techniques according to the type of techniques into three categories; traditional machine learning, graph mining, and deep learning.

There has been a large volume of work applying traditional machine learning methods based on data mining techniques for malware detection using different features. The first important work that used data mining techniques for malware detecting based on machine learning methods was done by [5]. They built different classifiers including instance-based learner, TFIDF, Naive-Bayes, support vector machines, decision tree, boosted Naive-Bayes, SVMs and boosted decision Tree. It showed that the best efficiency was reported by using the boosted decision tree J48 algorithm. This was followed by the work of [6]. They proposed a new method for identifying unknown malicious codes based on multiple classifiers and Dempster-Shafer theory. Their result demonstrated that the combination of classifiers outperformed individual classifier. In [7], the authors used hybrid feature retrieval (HFR) model to identify malicious executables. This model extracted three kinds of features from the executables by using assembly feature retrieval (AFR) algorithm and joins them into one list of feature called the hybrid feature set (HFS). They achieved a high level of accuracy in detecting malicious executables with a low rate of false positive. The idea of eDare system presented by [8]. They used eDare system in order to protect the users from eThreats and detect unknown threats. Their result showed that eDare system performed better than individual algorithms in terms of accuracy of prediction. In [9], the authors proposed Operation Code (OpCodes) as an extraction methods for malware detection and used n-grams of the OpCodes as the classification features. They showed that the accuracy rate exceeded 99% compared to byte sequence n-gram method, the small numbers of features led to enhance the performance of FS, and the DF was accurate for most of the top features. In the continuation of [9] work, [10] proposed to use active learning in the acquisition of unknown malicious in order to reduce the effort of labeling examples and keeping up the accuracy of classification. Their result demonstrated that the active learning method performed better than random learning method. The previous work of [10] was followed by [11] work. They used text categorization for the detection of unknown malicious based on four classification algorithms; Artificial Neural Networks (ANN), Decision Trees (DT), Naive Bayes (NB), and Support Vector Machines (SVM) with three kernel functions. Their experiment showed that the ANN and DT achieved a high rate of mean accuracies, more than 94%, and

low rate of false alarms and therefore the accuracy rate of 95% could be achieved by using a training set which has less than 20% of the malicious file. [12] introduced the idea of using an interpretable string based malware detection system (SBMDS) in order to identify the type of malware. Their results showed that SBMDS achieved more accuracy, efficiency, and scalability than anti-virus software and it performed better than data mining based detection systems and IMDS system. The Intelligent File Scoring System (IFSS) proposed by [13] in order to detect malware from the gray list. They showed that IFSS was more effectiveness and efficiency than NOD32 and Kaspersky. [14] proposed extracting variable length instruction sequences by using data mining techniques in order to find worms from programs. They built a decision tree, bagging and random forest as classification methods. Their experimental results showed that random forest performed well. In [15], the authors using API calls sequence feature for identifying malicious from clean files. They achieved a high rate of accuracy, exceeding 97%, in identify malware from cleanware. [16] used Anubis to extract the behavior of malware in a sandbox environment. They used five classifiers; k-Nearest Neighbors (KNN), Naive Bayes, J48 Decision Tree, Support Vector Machine (SVM), and Multilayer Perceptron Neural Network (MLP). Their study showed that the best performance was achieved by J48 Decision Tree with a precision of 97.3%, a true positive rate of 95.9%, an accuracy of 96.8%, and a false positive rate of 2.4%. The approach of semi-supervised learning was used by both [17] and [18]. They used a set of labelled and unlabeled file samples for malware detection. Their result showed that when supervised learning is used, the labelling efforts are lower while keeping up high accuracy rates. In [19], the authors proposed the technique of ensemble learning (EMPC) for automated classification of stream data. This technique used generalized, multipartition, and multi-chunk ensemble learning. They used synthetic data, botnet traffic, and malware dataset to evaluate this technique. Their experiment showed that EMPC technique was more efficient than other stream data classification techniques in term of detection accuracy and classification error and therefore it was more useful for the detection of intrusion. [20] introduced the idea of using file verdict system "Valkyrie" that use a model of semi-parametric classification in order to combine file content and file relations to improve the detection of malware. Their result showed that Valkyrie system performed better than other malware classification methods and anti-malware software in term of accuracy and efficiency. [21] and [22] focused on compiling static and dynamic analysis into a single test. In [21], the authors demonstrated that the merging of the three features of RF, DT and, IB1 into a single test enhanced the performance by 9% in term of accuracy of classification with improvement in the rate of false positive and the rate of a false negative. On the other hand, this method is less effective on the latest malware executables. [22] proposed OPEM approach as a hybrid malware detector. Their experiment showed that this approach enhanced the performance of classification. The idea of using GuardOL architecture was done by [23]. This

architecture used a frequency centralized model (FCM) to construct the feature for the purpose of learning the malicious behavioral manner from known malware samples. By using this architecture, they achieved high classification accuracy, high speed of detection, low consumption of power, and high flexibility. The main advantages of this architecture were the ability to detect 46% of malware during the first 30% of their execution, while this rate was increased to 79% during 100% of their execution and the rate of false positive was less than 3%. While machine learning methods based on data mining techniques and feature based on Windows API calls have been used in these studies [24] [25] [26] [27]. In [24] and [25], the authors demonstrated that the proposed system outperformed anti-virus software in term of accuracy, scalability, and efficiency. [26] and [27] proposed a Hierarchical associative classifier (HAC) and CIMDS systems respectively. They showed that the proposed systems were more effectiveness and efficiency than anti-virus scanners in term of detecting malware from the gray list.

Besides traditional machine learning techniques, a number of the authors have used graph mining methods for malware detection. For example, [28] used the implementation of HOLMES by combined the graph mining and the concept of analysis in order to get discriminative specifications to identify malware. By using HOLMES, the researchers have been able to increase the rate of malware detection. On the other hand, it needs a lot of labeled executables for supervised learning (classification) [18]. The work done by [29] used the Polonium algorithm for detecting malware. By using this algorithm, they achieved 85% rate of correctly detecting malware with one iteration while the rate of correctly identifying malware got better for an extra 2% with more iterations. [30] identified Malware based on instruction traces. They used the modified Ether malware analysis framework [16] to gather the traces dynamically. They used Graph kernels between instances to build the similarity matrix, 2 grams to estimate the transition probabilities in the Markov chain, and the measures of the Gaussian kernel and spectral kernel to construct the kernel matrix. They showed that the performance based on instruction traces was excellent but this method suffers from Complexity of Computation and therefore restricts its usage. Another work of malware detection was done by [31]. They used multiple features extractions; 2-gram byte sequences, control flow graph, disassembled OpCodes, dynamic instruction traces, miscellaneous file information, and system call traces to detect malware. They showed that the accuracy rate achieved 98.07%. The graphs induced by file relationships used by [32]. They showed that this method achieved a high degree of accuracy and scalability.

On the other hand, there are recent few papers have used deep learning methods for detecting malware. For example, [33] used a deep neural network model for malware detection. Their experimental results demonstrated that this model achieved a 95% detection rate at 0.1% false positive rate (FPR) and therefore it gave the best accuracy rate compared with the previous detection engines that use static features. The stacked

AutoEncoders (SAEs) model based on Windows API calls used by [34] for malware detection. This framework made up of feature extractor and deep learning based classifier. The results showed that this model achieved more effectiveness and efficiency. Also, it performed better than Artificial Neural Network, Support Vector Machine, Naive Bayes, and Decision Tree in term of malware detection.

Different from the previous work, based on a collection of 1473 worms and 1722 clean programs, resting on the analysis of Variable length instruction sequence, we attempt to detect internet worms by using machine learning methods based on data mining techniques.

TABLE I: Summary of Some related work to malware detection by using data mining techniques

Ref.	Features	Dataset size	Technique	Results
[5]	Binary n-grams	1971 benign and 1651 malicious	SVM, NB, DT, and their boosted versions	DT performed best
[24]	Windows API calls	12214 benign and 17366 malicious files	SVM, DT, NB, associative classification	93% accuracy and 97.2% detection rate
[6]	byte n-grams	423 benign and 450 malicious executable codes	Ensemble PNN	multiple classifiers achieved a good performance
[7]	binary n-gram, derived assembly features, and dynamic link library call	1967 benign files and 1920 malware	SVM, DT, NB, boosted DT, and BN	Hybrid Feature method achieved a high degree of accuracy
[8]	5-grams sequences and PE header	byte 7694 malicious and 22736 benign files	DT, ANN, and BN	96% accuracy, 93% true positive and 3% false positive rate
[25]	Windows API calls	12214 benign files and 17366 malicious	NB, SVM, and DT	93.07% accuracy rate
[9]	n-gram OpCode sequence	more than 30000 files	ANN, DT, NB, and SVM	99% accuracy rate
[10]	5-grams sequences	byte 1182 files	SVM	active learning outperformed random learning method
[11]	byte n-grams	7688 malware and 22735 benign files	ANN, DT, NB, and SVM	95% accuracy rate achieved by using a training set which has less than 20% of the malicious file
[12]	Interpretable strings	8320 benign files and 31518 malicious	DT, NB, SVM, Baggins	93.7% accuracy with Benign files and Backdoor, 88.30% with Spyware, 92.7% with Trojans, and 92.6% with Worms
[26]	Windows API calls	100000 files, 8000000 benign, and 8000000 malware	DT, NB, SVM, associative classification	96.3% precision rate
[27]	Windows API calls	15000 benign and 35000 malicious files	DT, NB, SVM, associative classification	88.2% detection and 67.6% accuracy rate
[13]	Windows API calls and interpretable strings	89626 files	associative classifier, SVM, ensemble of heterogeneous base-level classifiers	more than 91% detection rate
[14]	variable length instruction sequence	1444 worms and 1330 clean files	DT, random forest, and Bagging	random forest achieved an accuracy rate of 96% and false alarm rate of 3.8%
[15]	API call sequences executing in a virtual environment	1368 malware and 456 cleanware files	SVM, DT, random forest, instance-based classifier	97% accuracy rate
[16]	Behaviors extracted in sandbox environment	220 malware and 250 benign files	kNN, NB, DT, SVM, MLPNN	J48 DT achieved a precision of 97.3%, a true positive of 95.9%, an accuracy of 96.8%, and 2.4% false positive rate
[17]	n-gram distributions	17000 malicious programs	semi-supervised algorithm, collective learning approach	supervised learning achieved 90% accuracy rate with a low number of labeled files
[18]	n-gram distributions	1000 malicious and 1000 benign files	Semi-supervised algorithm, collective learning approach	supervised learning achieved 86% accuracy rate with a low number of labeled files

TABLE I – *Continued*

Ref.	Features	Dataset size	Technique	Results
[29]	file-to-machine relation graphs	48 million machine nodes and 903 million files	belief propagation	85% detection rate with one iteration and it got better for an extra 2% with more iterations
[30]	graphs constructed from DIT	1615 malware and 615 benign files	markov chain	96.4% accuracy rate
[19]	byte n-grams	105000 executables files	DT and ripper	EMPC achieved the lowest error rate
[20]	file content combining file relations	225830 benign and 434870 unknown files	semi-parametric classifier model	99.40% accuracy rate
[31]	2-gram byte sequences, disassembled OpCodes, DIT, miscellaneous files information, and SCT	780 malware and CFG, 776 benign files	instruction traces, SCT	98.07% accuracy rate
[32]	graphs induced by file relationships	more than four million containers with more than one executable file inside them	regression classifier	a high degree of accuracy and scalability
[21]	function length frequency, PSI, API function names, and API parameters	2398 malware and 541 benign files	DT, random forest, and instance-based classifier	integrating features achieved an accuracy rate of 97% with combined data set
[22]	Sequence of operational codes, SCO, and raised exceptions	1000 malware and 1000 benign files	DT, kNN, BN, and SVM	It enhanced the performance of classification
[33]	byte entropy histogram, string 2D histogram, PE import information, and PE metadata	81,910 benign files and 350,016 malware	DNN and Bayesian calibration model	95% detection rate at 0.1% false positive rate
[34]	Windows API calls	22500 malware, 5000 unknown, and 22500 benign files	DL Architecture using the stacked AutoEncoders	96% accuracy rate in the term of detecting malware
[23]	Resource-critical system call patterns	472 malware and 371 benign files	DT, NB, LR, SVM, sequential minimal optimization, RRL, and multilayer perceptron	46% detection at the first 30% of their execution, it was increased to 79% during 100% of their execution, and the rate of false positive was less than 3%

SVM:support vector machines, NB:Naive Bayes, DT:decision tree, PNN:probabilistic neural network, BN:Bayes networks, ANN:artificial neural networks, kNN:K-nearest neighbors, MLPNN:multilayer perceptron neural network, DIT:dynamically instruction traces, CFG:control flow graph, SCT:system call traces, SCO:system calls operations, LR:logistic regression, RRL:RIPPER rule learner

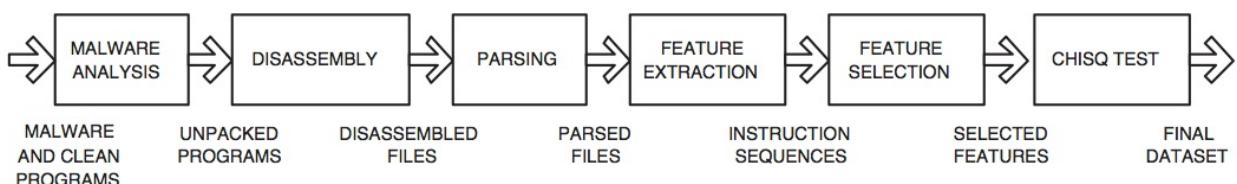


Fig. 1: Data preprocessing steps.

TABLE II: Packers/Compilers Analysis Details of Worms and Clean Programs

Packer/Compiler	Before Unpacking		After Unpacking	
	Worms	Cleans	Worms	Cleans
ASPack	79	2	1	0
Borland	118	39	258	45
FSG	31	1	3	0
Microsoft	350	937	649	976
Other Not Packed	205	597	234	601
Other Packed	104	24	135	28
PECompact	26	2	7	0
Unidentified	161	72	161	72
UPX	399	48	24	0
Total	1473	1722	1473	1722

TABLE III: Packers/Compilers Analysis Summary of Worms and Clean Programs

Packer/Compiler	Before Unpacking		After Unpacking	
	Worms	Cleans	Worms	Cleans
Not Packed	672	1573	1140	1622
Packed	640	77	171	28
Unidentified	161	72	162	72
Total	1473	1722	1473	1722

III. DATA PROCESSING

Our collection of 3195 Windows PE files consisted of 1473 worms and 1722 clean programs. The clean programs were obtained from a PC running Windows XP. These include small Windows applications such as calc, notepad, etc and other application programs running on the machine. A number of clean programs were also downloaded from [35] to get a representation of downloaded programs. The worms were downloaded from [36]. The dataset was thus consisted of a wide range of programs, created using different compilers and resulting in a sample set of uniform representation. Figure 1 displays the data processing steps.

A. Malware Analysis

We ran PEiD [37] on our data collection to detect compilers, common packers and cryptors, used to compile and/or modify the programs. Table II displays the distribution of different packers and compilers in the collection. Before further processing, packed programs were unpacked using specific unpackers such as UPX (with -d switch) [38], and generic unpackers such as Generic Unpacker Win32 [39] and VMUnpacker [40]. Table III displays the summary of packed, not packed and unidentified trojans and clean programs.

A number of clean programs were removed to keep an equal class distribution. The unidentified programs were also removed.

B. Disassembly

Binaries were disassembled to obtain a source code representation using Datarescues' IDA Pro [41]. Programs with disassembly errors were removed from the dataset.

```

inc    si
jb     short near ptr loc_171+1
ins    word ptr es:[di], dx
cmp    ah, [bx+si]
inc    di
popa
jz
dec
outsw
arpl  [bp+di+58h], bp
xor   dh, [bx+si]
xor   [bx+si+74h], al
jb    short near ptr loc_178+3

```

Fig. 2: Portion of the output of disassembled Netsky.A worm.

```

inc jb
ins cmp inc popa jz
dec arpl xor xor jb

```

Fig. 3: Instruction sequences extracted from the disassembled Netsky.A worm.

C. Feature Extraction

The core of the Feature Extraction Mechanism consisted of a parser that parsed the disassembled files to generate instruction sequences. A sequence is defined as instructions in succession until a conditional or unconditional branch instruction and/or a function boundary is reached. Instruction sequences thus obtained are of various lengths. We only considered the opcode and the operands were discarded from the analysis. Figure 2 shows a portion of the disassembly of the Netsky.A worm.

The parser was written in PHP and it translated the disassembly in figure 2 to instruction sequences. Figure 3 displays the output of the parser. Each row in the parsed output represented a single instruction sequence. For comparison purposes we also extracted non-overlapping fixed length instruction sequences from the disassembly. These fixed length sequences will be termed n-grams in this paper, where n is the length of the sequence. The value of n was varied from 2 to 10 in our experiments. A special case was also considered to experiment with long fixed size sequences. The n-gram size was set to 20 for this set.

D. Feature Selection

We used the *vector space model* to represent our datasets. In information retrieval, a vector space model defines documents as vectors (or points) in a multidimensional Euclidian space where the axes (dimensions) are represented by terms. Depending upon the type of vector components (coordinates), there are three basic versions of this representation: Boolean, term frequency (TF) and term frequency - inverse document frequency (TFIDF). [42]. In our case, the programs and instruction sequences mapped to documents and terms, respectively. Using these program vectors, we created a term-document matrix where programs were arranged in rows, while columns

represent the potential features (instruction sequences). For feature selection we used boolean representation of the matrix, where a 1 represented presence, while 0 represented absence, of an instruction sequence in a given program. Assume there are n programs p_1, p_2, \dots, p_n , and m instruction sequences s_1, s_2, \dots, s_m . Let n_{ij} be the number of times a sequence s_i was found in a program p_j . In the boolean representation a program p_j is represented as an m component vector, $p_j = p_j^1, p_j^2, \dots, p_j^m$,

$$p_j^i = \begin{cases} 0 & \text{if } n_{ij} = 0 \\ 1 & \text{if } n_{ij} > 0 \end{cases} \quad (1)$$

Using the boolean definition of p_j^i , let N_{ij} be the total number of times a sequence s_i was present in the program collection.

$$N_{ij} = \sum_{j=1}^n p_j^i \quad (2)$$

In order to be selected, a sequence s_i , must have its N_{ij} greater than a defined threshold. This threshold was set to 10% of the total number of the programs, as it is a common practice in data mining for defining unary variables.

$$N_{ij} > \frac{n}{10} \quad (3)$$

For each n-gram size, the term-document matrix was mostly a sparse matrix as most of the sequences were rare items. The unary variable removal step reduced the number of features to a fraction. The process was repeated for each n-gram size and the variable length set. Once the features were selected, a binary target variable was added to identify each program as worm or clean. We also added a heuristic flag variable indicating if the worm or clean program was originally found in a packed or an unpacked state.

E. Independence Test

A Chi-Square test of independence was performed for each feature to determine if a relationship exists between the feature and the target variable. For each feature a 2-way contingency table was created. Using a p-value of 0.01 for the test resulted in the removal of about 15%-20% of the features that did not show any statistically significant relationship with the target.

Table IV displays the features statistics for datasets generated for each n-gram size and the variable length instruction sequences.

F. Feature Reduction

For further processing, the term-document matrices for each n-gram size and the variable length sequences were transformed to the term frequency form from the boolean representation. After performing the feature selection and independence test, we applied two different feature reduction techniques to create two datasets. These techniques include random forest and principal component analysis. In addition to these two datasets we also kept the original dataset with

TABLE IV: Features Statistics

Feature Size	No of Features			
	Total	Distinct	Unary Removal	Ind. Test
2	16365620	6728	630	537
3	10910413	45867	1583	1334
4	8182810	156572	1823	1483
5	6546248	361715	1771	1466
Variable	5487145	509778	574	492
6	5455206	640863	1114	906
7	4675891	949233	458	363
8	4091405	1236822	112	71
9	3636804	1474690	37	27
10	3273124	1618698	20	15
20	1636562	1252860	2	0

TABLE V: Reduced Feature Sets

Feature Size	All Variables	RF Variables	PCA Variables
2	537	79	124
3	1334	124	148
4	1483	138	133
5	1466	138	120
Variable	492	89	164
6	906	148	164
7	363	99	200
8	71	35	68
9	27	15	27
10	15	10	15
20	0	0	0

all the features retained from the chi-square test output. In the rest of the paper, we will refer to this set as *All variables*.

1) *Random Forest*: Besides classification, random forest also gives the important variables used in the model. The importance is calculated as the mean decrease in accuracy or mean decrease in Gini index if the variable is removed from the model. We rejected the variables for which the mean decrease in accuracy was less than 10%. This dataset will be referred later in the paper as *RF variables*

2) *Principal Component Analysis*: PCA is a technique used to reduce multidimensional data sets to lower dimensions for analysis. PCA involves the calculation of the eigenvalues that represent the linear combination of original variables such that the lower order eigenvalues explain most of the variance in the data. We kept the variables that explained 95% variance in the dataset and rejected others. This dataset will be referred later in the paper as *PCA variables*

Table V displays the number of features in each dataset after applying the random forest and principal component analysis, for each feature size.

IV. EXPERIMENTS

Experiments were conducted on 30 different datasets. As explained in the previous section these were generated by using a combination of feature size and selection mechanism. Each dataset was partitioned into 70% training and 30% test data. Similar experiments showed best results with tree based models for the count data [43]. We built bagging and random forest models using R [44].

A. Bagging

Bagging or Bootstrap Aggregating is a meta-algorithm to improve classification and regression models in terms of accuracy and stability. Bagging generates multiple versions of a classifier and uses plurality vote to decide for the final class outcome among the versions. The multiple versions are created using bootstrap replications of the original dataset. Bagging can give substantial gains in accuracy by improving on the instability of individual classifiers. [45]

We used classification trees with 100 bootstrap replications in the Bagging model.

B. Random Forest

Random forest provides a degree of improvement over Bagging by minimizing correlation between classifiers in the ensemble. This is achieved by using bootstrapping to generate multiple versions of a classifier as in Bagging but employing only a random subset of the variables to split at each node, instead of all the variables as in Bagging. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost, but are more robust with respect to noise.[46]

We grew 100 classification trees in the Random forest model. Each random forest model was first tuned to obtain the optimal number of variables to be sampled at each node.

V. RESULTS

The test data contained new and unseen worms, that were not used in the training phase of the classifiers. The models performance was tested using this data. Confusion matrices were created for each classifier using the actual and predicted responses. The following four estimates define the members of the matrix.

True Positive (TP): Number of correctly identified malicious programs.

False Positive (FP): Number of wrongly identified benign programs.

True Negative (TN): Number of correctly identified benign programs.

False Negative (FN): Number of wrongly identified malicious programs.

The performance of each classifier was evaluated using the detection rate, false alarm rate and overall accuracy that can be defined as follows:

Detection Rate: Percentage of correctly identified malicious programs.

$$\text{DetectionRate} = \frac{TP}{TP+FN}$$

False Alarm Rate: Percentage of wrongly identified benign programs.

$$\text{FalseAlarmRate} = \frac{FP}{TN+FP}$$

Overall Accuracy: Percentage of correctly identified programs.

$$\text{OverallAccuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

TABLE VI: Experimental results for new and unseen worms

Classifier	Selection	Size	Det Rate	FP Rate	Acc
RF	All	var	94.3%	5.64%	94.24%
RF	RF	var	93.85%	7.6%	95.38%
RF	RF	5	93.04%	7.46%	93.56%
RF	RF	2	92.85%	5.74%	91.47%
RF	All	2	92.7%	6.85%	92.29%
Bag	RF	2	92.55%	6.65%	91.76%
Bag	All	var	92.5%	7.12%	92.12%
Bag	RF	var	92.35%	8.48%	93.23%
RF	RF	4	91.63%	7.45%	90.78%
RF	RF	8	91.63%	7.45%	90.78%
RF	RF	9	91.63%	7.45%	90.78%
Bag	RF	5	91.57%	8.28%	91.41%
RF	RF	3	91.51%	5.49%	88.63%
RF	PCA	4	91.38%	8.06%	90.8%
RF	PCA	5	91.38%	8.06%	90.8%
RF	PCA	6	91.38%	8.06%	90.8%
RF	PCA	7	91.38%	8.06%	90.8%
RF	PCA	8	91.38%	8.06%	90.8%
RF	PCA	9	91.38%	8.06%	90.8%
RF	All	3	91.21%	7.01%	89.5%
RF	All	6	91.15%	7.42%	89.86%
RF	All	5	90.92%	8.96%	90.8%
Bag	All	3	90.76%	7.01%	88.63%
Bag	PCA	5	90.62%	10.75%	92.02%
Bag	RF	3	90.61%	7.32%	88.63%
Bag	PCA	var	90.55%	9.5%	90.61%
RF	PCA	var	90.4%	8.31%	89.09%
Bag	All	2	90.31%	6.85%	87.71%
RF	All	4	90.28%	8.39%	89.05%
Bag	All	5	90.17%	10.15%	90.49%
RF	PCA	2	90.01%	10.91%	90.96%
RF	PCA	3	89.57%	8.23%	87.46%
RF	RF	7	89.29%	11.21%	89.81%
RF	All	7	89.29%	10.91%	89.49%
Bag	PCA	3	89.27%	8.23%	86.88%
RF	RF	6	88.7%	10%	87.54%
Bag	RF	4	88.64%	8.7%	86.17%
Bag	RF	8	88.64%	8.7%	86.17%
Bag	RF	9	88.64%	8.7%	86.17%
Bag	RF	6	88.55%	7.42%	84.93%
Bag	PCA	2	88.23%	13.27%	89.76%
Bag	All	6	87.94%	9.03%	85.22%
Bag	PCA	4	87.29%	13.76%	88.3%
Bag	PCA	6	87.29%	13.76%	88.3%
Bag	PCA	7	87.29%	13.76%	88.3%
Bag	PCA	8	87.29%	13.76%	88.3%
Bag	PCA	9	87.29%	13.76%	88.3%
Bag	RF	7	87.27%	12.12%	86.62%
Bag	All	7	87.11%	13.03%	87.26%
RF	All	8	81.13%	12.5%	75%
Bag	All	8	80.81%	11.84%	73.73%
RF	All	10	79.66%	12.35%	72.36%
RF	RF	10	79.66%	12.35%	72.36%
Bag	All	10	78.14%	15.14%	72%
RF	All	9	77.93%	22.64%	78.43%
Bag	RF	10	77.76%	15.94%	72%
RF	PCA	10	77.76%	14.34%	70.55%
Bag	PCA	10	77.57%	12.75%	68.73%
Bag	All	9	74.61%	27.92%	76.8%

Table VI displays the experimental results for each classifier, selection strategy and n-gram size combination.

Table VI indicates that the variable length instruction sequences resulted in a higher detection and a lower false positive rate than the fixed length instructions of various

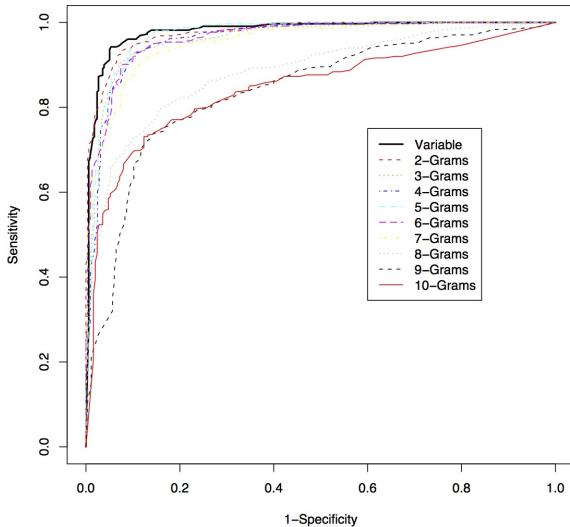


Fig. 4: ROC curves comparing random forest results for each n-gram size using all variables.

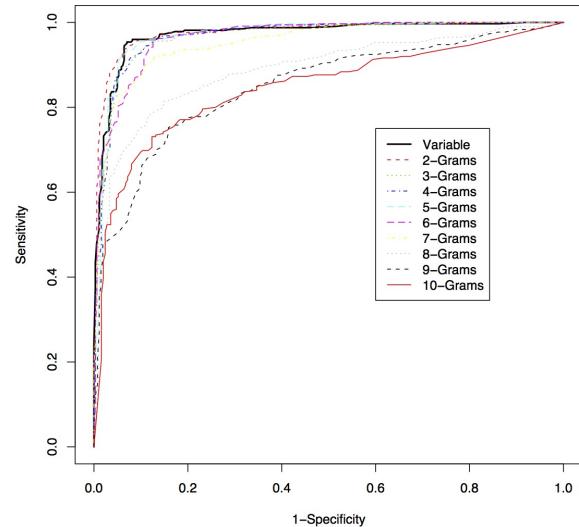


Fig. 5: ROC curves comparing random forest results for each n-gram size using random forest variable selection.

sizes. Combining the statistics of table IV with the results of table VI, it is also evident that variable length instruction sequences resulted in a better overall accuracy with a dataset of lesser dimensions, than the fixed length instruction sequences. Among the classifiers, random forest performed better than bagging which is endorsement of its superiority over bagging as claimed by [46].

Figures 4 - 11 display ROC curves comparing various combinations of classifiers, selection strategies and n-gram sizes. Figure 4 compares ROC curves for each n-gram size using random forest classifier on all the variables. Figure 5, compares similar curves on the random forest selection datasets, while figure 6, compares ROC curves on the PCA selection datasets. Figures 7, 8 and 9, compare similar ROC curves using bagging as the classification method. Figure 10 compares the results from each classifier on the random forest selected, variable length instruction sequences dataset. Another area of interest is to compare the different feature selection mechanisms used in these experiments. Figure 11 compares the ROC curves using random forest classifier on the variable length instruction sequences dataset using no selection, random forest selection and PCA selection.

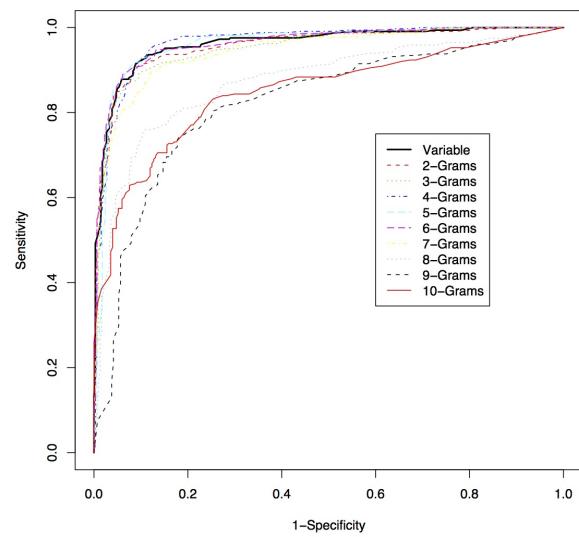


Fig. 6: ROC curves comparing random forest results for each n-gram size using PCA variable selection.

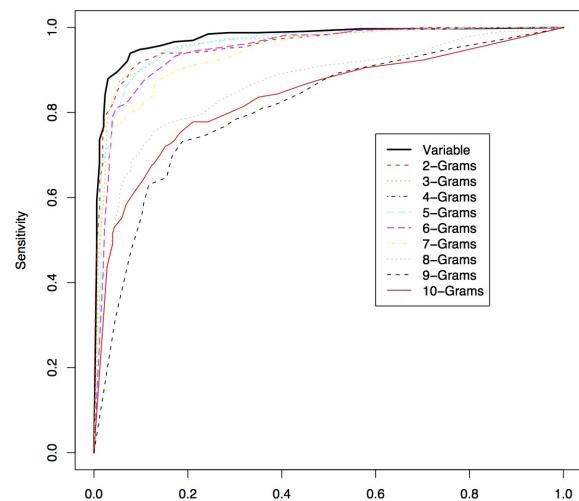


Fig. 7: ROC curves comparing bagging results for each n-gram size using all variables.

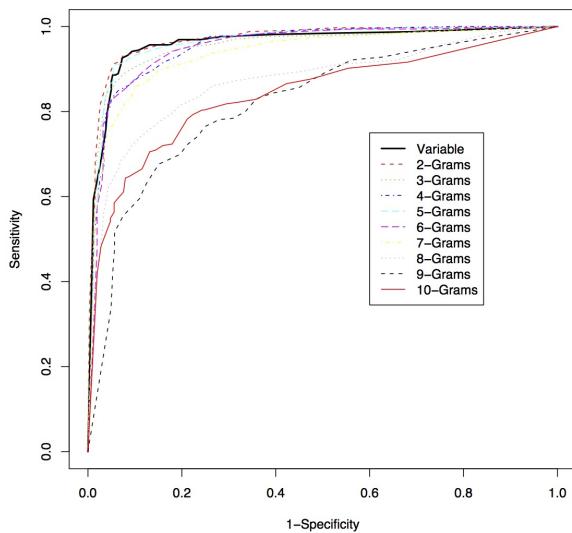


Fig. 8: ROC curves comparing bagging results for each n-gram size using random forest variable selection.

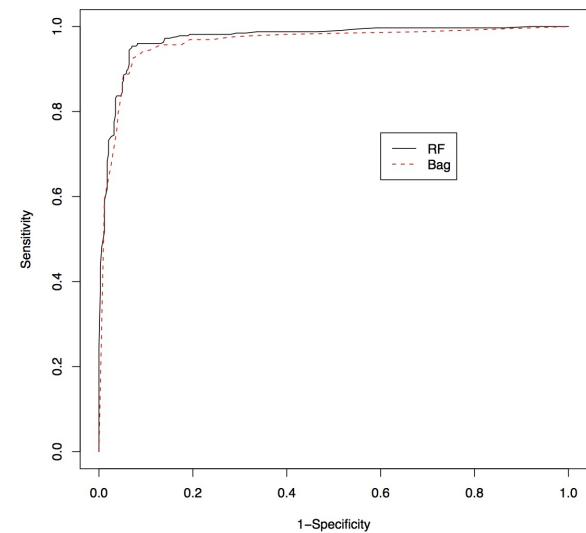


Fig. 10: ROC curves comparing random forest and bagging results using random forest variable selection.

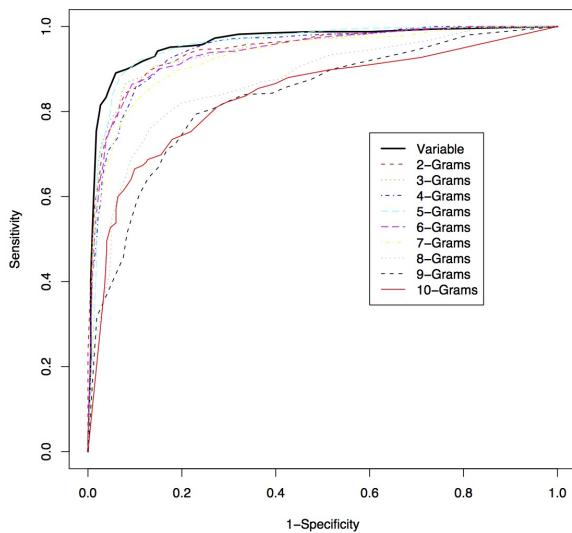


Fig. 9: ROC curves comparing bagging results for each n-gram size using PCA variable selection.

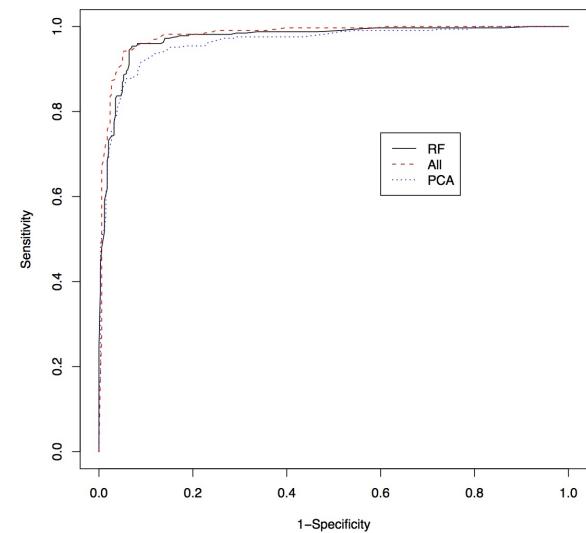


Fig. 11: ROC curves comparing the variable selection methods (all variables, random forest selection, PCA selection)

VI. CONCLUSIONS

In this paper we presented a data mining framework to detect worms using variable length instruction sequences and compared our features set to various fixed length instruction sequences. For a fair comparison, we extracted the variable length instruction sequences and the fixed length sequences from the same dataset. We used the vector space model to transform the disassembled data from the programs into a structured format by creating a term-document matrix. As the matrix was a sparse matrix, we used unary variable removal, a technique well known in data mining, to be the primary feature selection criteria. Our experiment demonstrated that the variable length instruction sequences resulted in a better overall accuracy using a smaller feature set. The result, is a simpler and robust statistical model with a higher detection and lower false positive rate. Another contribution of this experiment, is to compare different classifiers and different feature selection mechanisms. Our results displayed random forest to be the best classifier and the best feature selection mechanism.

REFERENCES

- [1] F. Cohen, "Computer viruses," Ph.D. dissertation, University of Southern California, 1985.
- [2] P. Szor, *The Art of Computer Virus Research and Defense*. New Jersey: Addison Wesley for Symantec Press, 2005.
- [3] Symantec, "Understanding heuristics: Symantec's bloodhound technology," Symantec Corporation, Tech. Rep., 1997.
- [4] M. Weber, M. Schmid, M. Schatz, and D. Geyer, "A toolkit for detecting and analyzing malicious software," in *Proceedings of the 18th Annual Computer Security Applications Conference*, 2002, p. 423.
- [5] J. Z. Kolter and M. A. Maloof, "Learning to detect malicious executables in the wild," in *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.
- [6] B. Zhang, J. Yin, J. Hao, D. Zhang, and S. Wang, "Malicious codes detection based on ensemble learning," *Autonomic and trusted computing*, pp. 468–477, 2007.
- [7] M. M. Masud, L. Khan, and B. Thuraisingham, "A scalable multi-level feature extraction technique to detect malicious executables," *Information Systems Frontiers*, vol. 10, no. 1, pp. 33–45, 2008.
- [8] Y. Elovici, A. Shabtai, R. Moskovitch, G. Tahan, and C. Glezer, "Applying machine learning techniques for detection of malicious code in network traffic," in *Annual Conference on Artificial Intelligence*. Springer, 2007, pp. 44–50.
- [9] R. Moskovitch, C. Feher, N. Tzachar, E. Berger, M. Gitelman, S. Dolev, and Y. Elovici, "Unknown malcode detection using opcode representation," *Intelligence and Security Informatics*, pp. 204–215, 2008.
- [10] R. Moskovitch, N. Nissim, and Y. Elovici, "Acquisition of malicious code using active learning," in *Proc. 2nd Intl Workshop on Privacy, Security, & Trust in KDD*, 2008.
- [11] R. Moskovitch, D. Stopel, C. Feher, N. Nissim, and Y. Elovici, "Unknown malcode detection via text categorization and the imbalance problem," in *Intelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on*. IEEE, 2008, pp. 156–161.
- [12] Y. Ye, L. Chen, D. Wang, T. Li, Q. Jiang, and M. Zhao, "Sbmds: an interpretable string based malware detection system using svm ensemble with bagging," *Journal in computer virology*, vol. 5, no. 4, pp. 283–293, 2009.
- [13] Y. Ye, T. Li, Q. Jiang, Z. Han, and L. Wan, "Intelligent file scoring system for malware detection from the gray list," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 1385–1394.
- [14] M. Siddiqui, M. C. Wang, and J. Lee, "Detecting internet worms using data mining techniques," *Journal of Systemics, Cybernetics and Informatics*, vol. 6, no. 6, pp. 48–53, 2009.
- [15] R. Tian, R. Islam, L. Batten, and S. Versteeg, "Differentiating malware from cleanware using behavioural analysis," in *Malicious and Unwanted Software (MALWARE), 2010 5th International Conference on*. IEEE, 2010, pp. 23–30.
- [16] I. Firdausi, A. Erwin, A. S. Nugroho *et al.*, "Analysis of machine learning techniques used in behavior-based malware detection," in *Advances in Computing, Control and Telecommunication Technologies (ACT), 2010 Second International Conference on*. IEEE, 2010, pp. 201–203.
- [17] I. Santos, C. Laorden, and P. G. Bringas, "Collective classification for unknown malware detection," in *Security and Cryptography (SECRYPT), 2011 Proceedings of the International Conference on*. IEEE, 2011, pp. 251–256.
- [18] I. Santos, J. Nieves, and P. G. Bringas, "Semi-supervised learning for unknown malware detection," in *DCAI*. Springer, 2011, pp. 415–422.
- [19] M. M. Masud, T. M. Al-Khateeb, K. W. Hamlen, J. Gao, L. Khan, J. Han, and B. Thuraisingham, "Cloud-based malware detection for evolving data streams," *ACM transactions on management information systems (TMIS)*, vol. 2, no. 3, p. 16, 2011.
- [20] Y. Ye, T. Li, S. Zhu, W. Zhuang, E. Tas, U. Gupta, and M. Abdulhayoglu, "Combining file content and file relations for cloud based malware detection," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 222–230.
- [21] R. Islam, R. Tian, L. M. Batten, and S. Versteeg, "Classification of malware based on integrated static and dynamic features," *Journal of Network and Computer Applications*, vol. 36, no. 2, pp. 646–656, 2013.
- [22] I. Santos, J. Devesa, F. Brezo, J. Nieves, and P. G. Bringas, "Open: A static-dynamic approach for machine-learning-based malware detection," in *International Joint Conference CISIS12-ICEUTE 12-SOCO 12 Special Sessions*. Springer, 2013, pp. 271–280.
- [23] S. Das, Y. Liu, W. Zhang, and M. Chandramohan, "Semantics-based online malware detection: towards efficient real-time protection against malware," *IEEE transactions on information forensics and security*, vol. 11, no. 2, pp. 289–302, 2016.
- [24] Y. Ye, D. Wang, T. Li, and D. Ye, "Imds: Intelligent malware detection system," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 1043–1047.
- [25] Y. Ye, D. Wang, T. Li, D. Ye, and Q. Jiang, "An intelligent pe-malware detection system based on association mining," *Journal in computer virology*, vol. 4, no. 4, pp. 323–334, 2008.
- [26] Y. Ye, T. Li, K. Huang, Q. Jiang, and Y. Chen, "Hierarchical associative classifier (hac) for malware detection from the large and imbalanced gray list," *Journal of Intelligent Information Systems*, vol. 35, no. 1, pp. 1–20, 2010.
- [27] Y. Ye, T. Li, Q. Jiang, and Y. Wang, "Cimds: adapting postprocessing techniques of associative classification for malware detection," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 3, pp. 298–307, 2010.
- [28] M. Fredrikson, S. Jha, M. Christodorescu, R. Sailer, and X. Yan, "Synthesizing near-optimal malware specifications from suspicious behaviors," in *Security and Privacy (SP), 2010 IEEE Symposium on*. IEEE, 2010, pp. 45–60.
- [29] D. H. Chau, C. Nachenberg, J. Wilhelm, A. Wright, and C. Faloutsos, "Polonium: Tera-scale graph mining for malware detection," in *Acm sigkdd conference on knowledge discovery and data mining*, 2010.
- [30] B. Anderson, D. Quist, J. Neil, C. Storlie, and T. Lane, "Graph-based malware detection using dynamic analysis," *Journal in computer Virology*, vol. 7, no. 4, pp. 247–258, 2011.
- [31] B. Anderson, C. Storlie, and T. Lane, "Improving malware classification: bridging the static/dynamic gap," in *Proceedings of the 5th ACM workshop on Security and artificial intelligence*. ACM, 2012, pp. 3–14.
- [32] N. Karampatziakis, J. W. Stokes, A. Thomas, and M. Marinescu, "Using file relationships in malware classification," in *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2012, pp. 1–20.
- [33] J. Saxe and K. Berlin, "Deep neural network based malware detection using two dimensional binary program features," in *Malicious and Unwanted Software (MALWARE), 2015 10th International Conference on*. IEEE, 2015, pp. 11–20.
- [34] W. Hardy, L. Chen, S. Hou, Y. Ye, and X. Li, "Dl4md: A deep learning framework for intelligent malware detection," in *Proceedings of the International Conference on Data Mining (DMIN)*. The Steering

- Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2016, p. 61.
- [35] "Download.com <http://www.download.com/>." [Online]. Available: <http://www.download.org/>
 - [36] "VX Heavens. <http://vx.netlux.org/>." [Online]. Available: <http://vx.netlux.org>
 - [37] "PEiD. <http://peid.has.it/>." [Online]. Available: <http://peid.has.it/>
 - [38] "UPX the Ultimate Packer for eXecutables. <http://www.exeinfo.go.pl/>." [Online]. Available: <http://upx.sourceforge.net/>
 - [39] "Generic Unpacker Win32. <http://www.exetools.com/unpackers.htm>." [Online]. Available: <http://www.exetools.com/unpackers.htm>
 - [40] "VMUnpacker. <http://dswlab.com/d3.html>." [Online]. Available: <http://dswlab.com/d3.html>
 - [41] "IDA Pro Disassembler. <http://www.datarescue.com/idabase/index.htm>." [Online]. Available: <http://www.datarescue.com/idabase/index.htm>
 - [42] Z. Markov and D. T. Larose, *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*. Wiley-Interscience, 2007.
 - [43] M. Siddiqui, M. C. Wang, and J. Lee, "Data mining methods for malware detection using instruction sequences," in *Proceedings of Artificial Intelligence and Applications, AIA 2008*. ACTA Press, 2008.
 - [44] "The r project for statistical computing <http://www.r-project.org/>." [Online]. Available: <http://www.r-project.org/>
 - [45] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996. [Online]. Available: citeseer.ist.psu.edu/breiman96bagging.html
 - [46] ———, "Random forests." *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: citeseer.ist.psu.edu/breiman01random.html

Tuning of Canny Image Edge Detection

Jamil A. M. Saif

Hodeidah University, Hodeidah , Yemen

Currently : University of Bisha, Bisha KSA

Jamil_alabssi@yahoo.com

Abstract – Image edge detection is the process of detection the pixel's intensity change between two adjacent regions in an image, but this considered to be a challenging issue due to noises existence as well as the type of image itself, for example in endoscopic images it is hardly ever to distinguish between regions by a physician doctor who may interpret the image content, for that reason image edge detection used as a method for image segmentation. The Canny edge detector is one of the most salient operators of image edge detection. So in this paper the problem of tuning canny image edge detection is addressed , and the affect of the threshold and sigma parameters of canny algorithm are tested and analyzed, resulting in determination of the optimal ones for natural as well as medical images, which finally lead to more accurate edge images.

Index Terms – image edge detection, Canny edge detection, canny threshold.

1. INTRODUCTION

Edge detection is considered to be one of the most challenging issues relevant to image processing and image analysis[1,3,4,8]. Its quality and precision ultimately affect the results of object recognition. Several application[3,9,11] are basically are dependable on the output of image edge detectors. Image edge detection is defined as the transformation of gray scale image into a binary image, that rely on sudden changes of pixel's intensity value, such techniques are named boundary based methods[2,4,6,13]. In such methods abrupt changes are searched to produce a binary image which represents the edges and the background of instigated image. From the edges many useful features can be extracted, thus these features are exploited by upper level computer vision algorithms[1,3,12,15]. Edge detection plays a crucial role in several applications such as object detection, recognition medical diagnosis and many others. . Image edge detection is an open and promising field of research, many researches deal with a single based derivative edge detectors [2,4,5]. In this paper a Canny detector will be applied for our experiments to investigate its performances with changing the parameters of threshold and standard deviation.

The contents of this paper is organized as follows: in section 2 Canny algorithm is given in details, experimental results are presented and discussed in section 3 the, conclusion and future work is outlined in section 4.

2. The Implemented algorithms

A variety of image edge detection algorithms[5,6,11,13] have been applied for the process of image edge detection. Edge images are produced as an output of such algorithms, that image is simply a binary image representing the objects within the original image and its background. Such algorithms search for a local change of pixel's intensity, and as a result produce edge images. Among such algorithms are Sobel, Prewitt, Roberts, LoG and Canny edge detection algorithms , in our research we focused on Canny Edge detector and the tuning of its parameters that affect the performance of this algorithm.

2.1. Canny Edge Detector

Canny edge detector is one of several of image edge detectors that have been used[9,13,17,18], but Canny edge detector is considered to be more efficient, it finds edges with more accuracy thanks to minimal of signal to noise ratio. Canny edge detector is characterized due to the following features:

- I. Signal to noise ratio of the gradient is maximized, leading to High quality of edge detection.
- II. Good localization, localized edges should be with maximal accuracy to real ones
- III. Minimal responses, the detector should give a single edge and eliminate false edge caused by noises.

The description of Canny algorithm [6,17,18] is given as follows:

1. Image smoothing by convolution with Gaussian filter as a result noises are removed.
2. Magnitude and angle of Gradient Calculation for each pixel of the smoothed image in the horizontal and vertical directions[16,17]
3. Thinning image by application of non- maximum suppression of image's pixels, that suppress all pixels except the ones with local maxims, as a result thin edges are preserved.

4. Hysteresis, two dynamic thresholds are used upper and lower represented by th and tl respectively, these are local thresholds and dependable on the local content within the image. Pixels with gradient values greater than th regarded as edge, and the ones that are smaller than tl are neglected, the ones between both threshold are considered as edges if they are connected to pixels with values higher than th otherwise they are rejected.

3. experimental results

The paper presents the effect of tuning canny parameters (threshold and sigma) to show how the performance of canny algorithms depends on these parameters.

Canny Edge detector algorithm computes pixel's gradient values by using a Gaussian filter, for that two threshold are applied upper and lower, to detect strong and weak edges respectively, the ones between are added to edges only if they are connected to strong edges. The accuracy of this algorithm guarantees to detect true weak edges[4,14,16].

Several medical and natural images are tested to show the Canny image edge detectors with a variety of parameters combinations(threshold and sigma). Only a sample of tested images are shown and the impact of canny parameters as shown in figures 1,2,3and 4.



Original image



threshold=0.6



threshold=0.5



threshold=0.4



threshold=0.3



threshold=0.2



threshold=0.18



threshold=0.16



threshold=0.15



threshold=0.14

Fig.1 the impact of canny threshold with fix sigma=1.414(for endoscopic image)



Sigma=1.4



sigma=1.6



sigma=1.8



sigma=2



sigma=2.5



sigma=3

original image

threshold=0.2



sigma=4



sigma=6



threshold=0.16



threshold=0.14



sigma=8



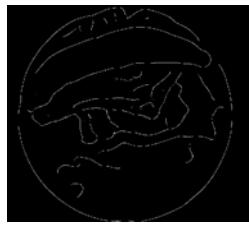
sigma=10



sigma=1.6



sigma=2



sigma=12



sigma=14



sigma=4



sigma=6

Fig.2 the impact of canny sigma with fix threshold=0.14 (for endoscopic image)



4. Conclusion

In this paper a heuristic approach has been performed giving the best possible of Canny images edge, as well as showing the effects of threshold and sigma variation on the output of images edge.

As it shown by the result Canny edge detector is more dependable on threshold which is a vector of two values lower and higher thresholds that are relative to detect

weak and strong edges respectively, and somehow dependable on sigma which determine the size of Gaussian filter. The optimal ones are varies one type of image to another one.

As it occurs from the results the most optimal threshold is around 0.14, and the sigma between 1.4 and 2, but for endoscopic image it is more fluctuate.

For future work an artificial approaches for adjusting the Canny parameters are highly recommended.

REFERENCES

- [1] T. Acharya and A. K. Ray. Image Processing Principles and Applications, John Wiley & Sons, Inc.,2005.
- [2] J. K. Anil. Fundamentals of digital image processing, Prentice Hall, April, 2004.
- [3] B. Chanda .. D. D. Majumder. Digital Image Processing and Analysis, Prentice Hall, 2003.
- [4] R. C. Gonzalez and R. E. Woods. Digital Image Processing, Second Edition, Prentice Hall, 2002.
- [5] J. A. Madhuri. Digital Image processing, Prentice Hall, 2006.
- [6] R. Maini and H. Aggarwal. Study and Comparison of Various Image Edge Detection Techniques. International Journal of Image processing
- [7] W. Malina, S. Ablameyko , W. Pawlak. Fundamental Methods of Digital Image Processing. (In Polish), 2002.
- [8] J. Saif and A. Moharram, Edge and Regin Based Image Segmentation, Journal of Computer and Information Technology of Hodeidah University, Hodeidah, Yemen, 2011.
- [9] R. Saini, M. Dutta and R. Kumar. A COMPARATIVE Study of Several Image Segmentation Techniques, Journal of Information and Operations Management, 2012.
- [10] E. A. Savakis. Adaptive Document Image Thresholding Using Foreground and Background clustering, published in Proceeding of International Conference on Image Processing ICIP, 98.
- [11] M. Sonka, V. Hlavac and R. Boyle, Image Processing, Analysis and Machine Vision. Thomson, 2008.
- [12] L. Spirkovsk. A Summary of Image Segmentation Techniques, Ames Research Center, Moffett Field, California, 1993.
- [13] R. Szeliski. Computer Vision - Algorithms and Application, Springer, 2011.
- [14] P. Thakare. A Study of Image Segmentation and Edge Detection Techniques, International Journal on Computer Science and Engineering(IJCSE), Feb. 2011.
- [15] S. E. Umbaugh. Computer Vision and Image Processing: A Practical Approach Using CVIP tools, Prentice Hall, 1998.
- [16] <http://suraj.lums.edu.pk/~cs436a02/CannyImplementation.htm>
- [17] http://www.codeproject.com/KB/cs/Canny_Edge_Detection.aspx
- [18] http://www.cvmt.dk/education/teaching/f09/VGIS8/AIP/canny_09gr820.pdf

The Necessity of Developing a Standard for Exchanging a Chain of Custody of Digital Evidence Data

A DEMF STORY

Jasmin Cosic

IT Section of Police Administration, MoI
of Una-sana canton
502.V.bbr, Bihać, Bosnia and
Herzegovina

jasmin.cosic@mupusk.gov.ba

Miroslav Baca

Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42 000 Varaždin
Croatia

mbaca@foi.hr

Petra Grd

Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42 000 Varaždin
Croatia

petra.grd@foi.hr

Abstract— Today there is no criminal investigation that does not contain a digital dimension. A large number of criminal offenses, whether official investigations conducted by judicial bodies or corporate investigations, contain digital evidence, which in most investigations is key to the identification of perpetrators. Since the cyber space is undefined, it has no owner, place, time dimension, it does not belong to anyone, very often such evidence must be exchanged between the various subjects involved in the investigation. However, in addition to the exchange of digital evidence, it is also necessary to exchange the so-called “5ws&1h data” or metadata that are key to chain of custody identification. This is necessary because of the large number of factors that can influence the evidence and undermine the integrity of digital evidence, after which the evidence will not be accepted by the court. The need for the standardization of metadata exchange procedures and processes that ensure the chain of custody has been imposed as a necessity and through the realization of DEMF, as a possible solution.

I. INTRODUCTION

Every digital investigation nowadays involves digital evidence collected from different sources, in different ways, by various institutions and people employed in these institutions. Investigations can be criminal, official, can be internal and corporate within the organization (firms), but what is common to all is digital evidence. Given that the cyber space is undefined, often the site of execution is the Internet in one country, the affected is located in another country and the perpetrator is geographically on the other end of the world. This requires a special, multi-functional approach, as well as the procedures to be followed in this process. The particular problem here is the absence of procedures in many countries which exchange digital evidence, for example the SE European countries.

This paper will outline the usage of Digital Evidence Management Framework (DEMF), the framework outlined in previous researches [1] [2] [3] [4] that allows guidance and

proving of 5ws&1h, or chain of custody at any time in any phase of digital forensic investigation. At every moment it is known who, what, when, where, why and how handled digital evidence. The need to set standards will also be emphasized, for example, the standardization of metadata exchange along with digital evidence, which DEMF provides. In the event that institutions or agencies from one country have a need to exchange digital evidence with another institution or agency in another country, along with digital evidence a .demf file will be provided that will contain all the metadata of the complete digital life cycle that will ultimately determine and prove that the digital evidence has not changed or that its integrity has not been violated in any phase of the digital forensic investigation.

Likewise, what DEMF allows and where its strength lies on the other hand, is not only in the ability to exchange metadata 5ws&1h that provides evidence of digital evidence integrity, but also the ability to exchange data in the so-called “container”.

II. RELATED WORK

There are many papers on standardization and interoperability of digital evidence.

One of the ways in which scientists tried to solve the problem of preserving the chain of custody was through creating different formats for storing digital evidence. One of the efforts of authors who are actively involved in the problem of standardizing the format of digital evidence storage as a prerequisite for the exchange of digital evidence, forensic tools and so on was also an attempt by members of the DFRWS working group. The recommended framework was built using the RDF (Resource Description Framework) as the most common and most affordable data presentation format and ontology for describing the vocabulary relevant to this

data. The methods used by authors in their paper are ontological approaches (ontologies), modeling, the Unified Modeling Language (UML), XML (eXtensible Modeling Language) and RDF (Resources Description Framework) [5]. It is emphasized that the ontology of the Philip Turner Digital Evidence Bag was applied, and there are not many papers where ontology has been applied in the field of digital forensics and computer crime. The paper did not solve the problem of the digital evidence chain of custody, and the ontology was used only to define the vocabulary used in defining this format.

Digital Evidence Bag (DEB) [6] [7] represents a universal container for storing digital evidence collected from any source. It ensures that not only data (potential digital evidence) can be stored but also the source of evidence and maintain continuity during the investigation process. In other words, DEB is part of the software that can store data from any forensic tool that is running. The DEB consists of the tag file, the index file and the contents file. The metadata file contains the metadata of the digital evidence (the name of the organization, the name and surname of the person who collected the evidence, date and time of collection, ID number, hash function calculation). The index file contains information about the corresponding content file (path to the data carrier, file name, time stamp). The content file is the actual digital evidence (image, video, text file, and so on). There is no detailed elaboration of this concept, and according to some authors [5] DEB can be any archive (tar, zip, etc.) that contains these files. The most common and publicly available formats for digital evidence storage are AFF, Raw, DEB, Expert Witness, Gfzip, ProDiscovery, EnCase, and SMART Expert Witness [8]. Some of these tools have built-in digital evidence integrity mechanisms through the combination of MD5 and CRC while only the SMART default format has cryptographic signature support.

Compared to other formats, DEMF has user-defined metadata, strong summary functions (MD5, SHA256/384/512), AES256 encryption support, as well as all metadata required to demonstrate the chain of evidence integrity. It also enables the original evidence (first copy) storage in the container itself.

Table 1 shows the most popular formats for storing digital evidence with the features important for chain of evidence and the protection of digital evidence integrity. It is evident that no format contains metadata that would answer the question 5ws&1h and the way to maintain a digital chain of evidence.

In comparison, for the first time, DEMF, as a possible solution that offers complete control of 5ws&1h, is presented.

TABLE I
MATRIX OF AVAILABLE FORMATS

Format name	Forensic tools support	Has metadata	Methods for DE integrity	Chain of custody metadata
Raw	Any forensic tool	-	-	-
AFF	AFF tools	Case	MD5	User

		number, investigator ID, Evidence number, SN, Time, Notes		defined (any combination of name/values)
AFF4	LibAFF4 AFF4 tools	Case number, investigator ID, Evidence number, SN, Time, Notes	MD5	User defined (any combination of name/values)
DEB	DEB viewer, imager, cmd wrapper	Agency name, investigator ID, Notes, Locations, Date and Time, host ID, etc.	Hash, encryption	Date and Time, application ID, Signature, hosts ID, Access to components
EnCase	EnCase, FTK, SMART, X-Ways, AFF	Case number, Evidence number, Notes, Time	MD5, CRC-32	-
GfZip	GFZ Tools (lib)	Like in AFF	SHA1, MD5, SHA256, X509, cryptographic signature	Cryptographic signed metadata
ProDiscover	ProDiscover	Disk image number, investigator ID, Time, System-time	MD5, SHA1, SHA256, digital signature	NO
SMART Expert Witness Comp	SMART, FTK imager	Case number, Investigator ID, Evidence number, Notes, Time	MD5 CRC-32	NO
DEM	DEM, DEMF Viewer	User defined metadata (Institution name, Case number, Summary, Date)	MD5 SHA256 SHA384 SHA512 AES256 encryption	5WS&1H Hash of evidence, geo-data, time-stamp, person ID,

		and Time, etc.	Court order number, Legislati on secured with AES256 encrypti on

Casey and other [9] developed CASE and DFAX standard to presents an open community-developed specification language called Cyber-investigation Analysis Standard Expression (CASE). To further promote a common structure, CASE aligns with and extends the Unified Cyber Ontology (UCO) construct, which provides a format for representing information in all cyber domains. This ontology abstracts objects and concepts that are not CASE-specific, so that they can be used across other cyber disciplines that may extend UCO. His work is a rational evolution of the Digital Forensic Analysis eXpression (DFAX) for representing digital forensic information and provenance. CASE is more flexible than DFAX and can be utilized in any context, including criminal, corporate and intelligence. [9][10]

For the Digital Forensic Community, a unique framework is needed - to help that the chain of digital evidence can be treated the same way around the world. Nowadays, common practice is (especially in the case of terrorist attacks) that some criminal activities are agreed on in one state on one side of the world and are committed in another country on the other side of the world, and the means of execution (a digital device) is at the very third end of the world. The question they ask is where is digital evidence, and who, when, where, why and how handles it, did the evidence remain unchanged during this process? How can the metadata be shared along with the digital evidence themselves to the interested parties involved in the investigations?

III. PROPOSED FRAMEWORK

DEMF is an almost 10 years old idea, proposed in conceptual model of Cosic and Baca in their early work [11][12].

The DEMF not only allows recording and managing the chain of evidence at all stages of the digital forensic investigation, but also ensures the integrity of the digital evidence itself. It also enables packing of all 5ws&1h data together with digital evidence and then secure protection with the help of powerful AES256 encryption. The model can be applied and used in digital investigations when we want to prove that the proof was not altered and that it is known at any time who, when, where, where, why and how came into contact with digital evidence throughout the life cycle.

Later, DEMF was realized as an application created in Java, and a case study with specific data was made - a test scenario. The power of DEMF is not just a chain of custody and

assurance of metadata integrity, but also the possibility of preserving the whole case (digital evidence and their metadata), but also the chain of evidence meta data in one container. The so-called container or .demf file is additionally secured with AES256 encryption, allowing full protection. This practically means that the exchange of evidence between the participants in the digital forensic investigation process would also exchange digital evidence themselves, their metadata, as well as the metadata needed to prove the chain of custody. On the other hand, DEMF's strength lies in the fact that this tool has integrated fully functional forensic tools features, because the tool reads data from digital evidence with the help of built-in libraries. The example of using DEMF can be seen in Figures 1, 2 and 3.

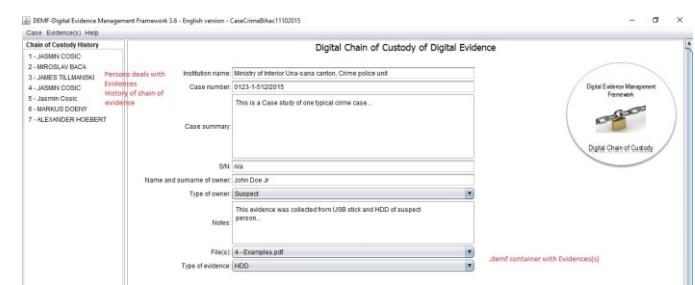


Figure 1 Example 1 of one real case

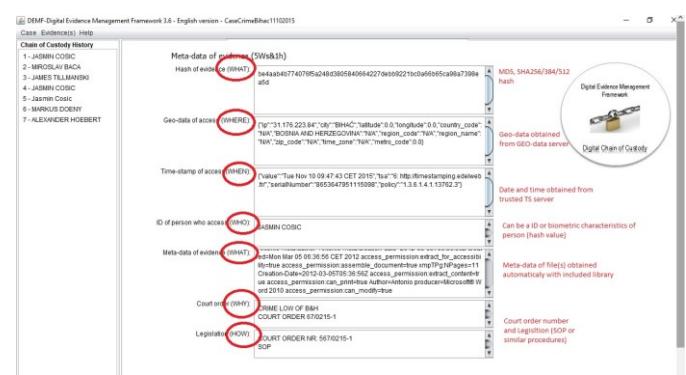


Figure 2 Example 2 of one real case

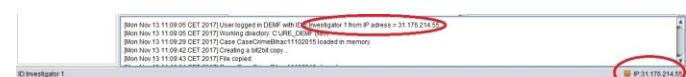


Figure 3 Example 3 of one real case with DEMF in action

The *.log file* that represents DEMF in action of collecting meta data can be seen below:

```
[Mon Nov 13 11:09:05 CET 2017] User logged in DEMF with ID = Investigator 1 from IP address = 31.176.214.55
[Mon Nov 13 11:09:05 CET 2017] Working directory: C:\JRE_DEMF (x86)
[Mon Nov 13 11:11:15 CET 2017] CaseDEMFI_256K2017 created in memory.
```

```
[Mon Nov 13 11:11:38 CET 2017] Getting a hash of file..  
[Mon Nov 13 11:11:38 CET 2017] Hash file is retrieved.  
[Mon Nov 13 11:11:38 CET 2017] Reading a meta-data..  
[Mon Nov 13 11:11:40 CET 2017] Meta-data of file is retrieved =  
Name=Slika1.jpg X-Parsed-By=org.apache.tika.parser.DefaultParser X-  
Parsed-By=org.apache.tika.parser.jpeg.JpegParser Resolution Units=inch File  
Modified Date=Mon Nov 13 11:11:40 CET 2017 Compression  
Type=Baseline Data Precision=8 bits Number of Components=3  
tiff:ImageLength=1024 Component 2=Cb component: Quantization table 1,  
Sampling factors 1 horiz/1 vert Component 1=Y component: Quantization  
table 0, Sampling factors 2 horiz/2 vert Image Height=1024 pixels X  
Resolution=96 dots Image Width=1280 pixels File Size=239026 bytes  
Component 3=Cr component: Quantization table 1, Sampling factors 1  
horiz/1 vert File Name=apache-tika-8052293475143496132.tmp  
tiff:BitsPerSample=8 tiff:ImageWidth=1280 Content-Type=image/jpeg Y  
Resolution=96 dots  
[Mon Nov 13 11:11:40 CET 2017] Getting a time-stamp..  
[Mon Nov 13 11:12:45 CET 2017] Time-stamp is retrieved.  
[Mon Nov 13 11:12:45 CET 2017] Getting a geo-data from server..  
[Mon Nov 13 11:12:45 CET 2017] Geo-data successfully retrieved from  
server.  
[Mon Nov 13 11:12:45 CET 2017] File  
C:\DEMF\CaseCrimeBihac11102015\Photo1.jpg successfully assigned to  
DEMF.
```

IV. CONCLUSION

The exchange of digital evidence between different institutions is necessary for a number of reasons. Given the undefined cyber space, a number of agencies/institutions are often involved in digital forensic investigations. The interoperability of digital evidence is necessary but it is also necessary to exchange metadata and chain of evidence. Today, dealing with chain of evidence in the world is often handled manually, while there are systems and solutions that offer electronic recording. The DEMF framework, in addition to conducting a chain of evidence, ensures the integrity and inviolability of digital evidence, as well as the proving of 5ws&1h which is the strict procedure on which the courts insist (Daubert principle). Every moment, we must know what, who, where, when, why and how he has accessed digital evidence. Given that, a large number of participants is involved in this process - from investigators, court attorneys, court experts, prosecutors, judges, police officers, bystanders and similar, there will be a large amount of data, and in every single step of the chain all data must be collected.

DEMF as the proposed framework, but also a finished solution, offers complete control of digital evidence management. In addition to metadata exchange, it also offers control of the digital evidence itself and proves its inviolability and integrity, which is the most important in the forensic investigation.

ACKNOWLEDGMENT

Research and development of DEMF in the first phase was supported by the Centre for biometrics, forensics and privacy, Faculty of Organization and Informatics, University of Zagreb.

REFERENCES

- [1] J. Cosic and M. Baca, "Do we have full control over integrity in digital evidence life cycle?," in *Proceedings of the International Conference on Information Technology Interfaces, ITI*, 2010, pp. 429–434.
- [2] J. Cosic and M. Baca, "Leveraging DEMF to Ensure and Represent 5ws&1h in Digital Forensic Domain," *Int. J. Comput. Sci. Inf. Secur.*, vol. Vol. 13, N, 2015.
- [3] M. Baca and J. Cosic, "Using DEMF in Process of Collecting Volatile Digital Evidence," in *The 39th International ICT Convention – MIPRO 2016 (IEEE), At Opatija, Croatia*, 2015.
- [4] J. Cosic, "Formal Acceptability of Digital Evidence," in *Multimedia Forensics and Security, SPRINGER*, Intelligen., SPRINGER, AG, 2017, pp. 327–348.
- [5] B. Schatz and A. Clark, "An open architecture for digital evidence integration," in *AusCERT Asia Pacific Information Technology Security Conference*, 2006, no. May, pp. 21–26.
- [6] B. Pladna, "Computer Forensics Procedures , Tools , and Digital Evidence Bags : What They Are and Who Should Use Them," in *Computer Forensics Procedures, Tools, and Digital Evidence Bags* 3, 2009, pp. 1–15.
- [7] P. Turner, "Applying a Forensic Approach to Incident Response, Network Investigation and System Administration using Digital Evidence Bags," *Digit. Investig.*, vol. 4 (1), no. 1, pp. 30–35, 2007.
- [8] DFRWS and CDESC Working Group, "Survey of Disk Image Storage Formats," 2006.
- [9] E. Casey, S. Barnum, R. Griffith, J. Snyder, H. van Beek, and A. Nelson, "Advancing coordinated cyber-investigations and tool interoperability using a community developed specification language," *Digit. Investig.*, pp. 1–32, 2017.
- [10] E. Casey, G. Back, and S. Barnum, "Leveraging CybOX™ to standardize representation and exchange of digital forensic information," *Digit. Investig.*, vol. 12 [1] E. C, pp. S102–S110, 2015.
- [11] J. Čosić and M. Baća, "(Im)proving chain of custody and digital evidence integrity with time stamp," in *MIPRO 2010 - 33rd International Convention on Information and Communication Technology, Electronics and Microelectronics, Proceedings*, 2010, pp. 1226–1230.
- [12] J. Cosic, Z. Cosic, and M. Baća, "Chain of Digital Evidence Based Model of Digital Forensic Investigation Process," *Int. J. Comput. Sci. Inf. Secur.*, vol. 9, no. 8, pp. 18–24, 2011.

Survey of Research Challenges in Cyber Physical Systems

Swati Nikam , Research Scholar, DIT, Pune, India

Swatinikam3@gmail.com

Dr. Rajesh Ingle, Sr. Member, IEEE

Abstract- In Cyber Physical Systems(CPS) there is a tight integration between cyber and physical world where individually the characteristics of both are different . So CPS is good example of heterogeneity in different aspects like the components used, data, the communication methodology which they adopt etc. So various research challenges are seen in the domain of CPS. We have discussed few research challenges such as Service Composition, Resource Provisioning and Autonomics. It also talks about various architectures which have been used by researchers. This paper focuses on different research challenges and the scope for research in those areas.

I. INTRODUCTION

CYBER PHYSICAL SYSTEMS is the recent topic in the field of information technology. Cyber-physical systems (CPS) are currently of interest everywhere including academia, industry, and government because CPS has a great potency to change the present scenario. The term Cyber Physical System was coined by Helen Gill at the National Science Foundation in the United States in 2006. Most recently, it has emerged as a promising direction to enrich the human-to-human, human-to-object, and object-to-object interactions in the physical world as well as in the virtual worlds[1]. CPS is defined as sensing, communication and processing platforms, deeply embedded in physical processes which provide real-time monitoring and actuation services. CPS is recognized as enabling technology as they enable numerous innovative applications[2]. It is evolving from various domains like wireless sensor network, control theory, embedded systems, distributed systems and many more. Cyber Physical System is the system that bridge the cyber world of computing and communication with the physical world. There is huge diversity of application domain like transportation, defense, industrial automation, health care and biomedical, critical infrastructure monitoring, agriculture and many more[3],[4],[5].

This paper is organized as follows. Section II highlights general research challenges in CPS. Section III is an overview of research work done along with research gaps in specific research areas of CPS such as Service Composition , Resource Provisioning and Autonomics. Section IV is about research problem found followed by concluding remarks in section V.

. II. RESEARCH CHALLENGES IN CPS

The interesting thing about CPS is though it covers a wide range of domains in the area of information technology but none of the them can be readily accepted in the context of CPS[6],[7].[8]. It needs a different consideration in the context of CPS. Although the additional research challenges will be present as per the application domain but common research challenges which are cutting across many domains can be listed as follows[9],[10],[11].

a) CPS Composition : Dynamic service composition is difficult because of joint interaction between physical and cyber processes on which heterogeneity of devices, communication technology affects a lot.

- b) Robustness and Safety of CPS : As CPS is application oriented so these functional and non functional characteristics needs to be satisfied.
- c) Security of CPS : As the CPS has a range of users, devices, computation at various levels and various communication technologies are associated with that, so securing the CPS is needed.
- d) Computational Abstractions : Various Physical characteristics should be captured in a compostable manner in a programming abstractions.
- e) Architectures: As CPS architecture must be consistent at a meta level and capture a variety of physical information.
- f) Verification, Validation and Certification of CPS : There is a gap between formal methods and tools which are used in practice which needs to be bridged.
- g) Autonomics : As autonomies is the desired requirement of any CPS, so achieving autonomies is the necessary thing as nobody wants CPS to crash at any time. Hence autonomies of self management is very much required to design a self sustaining CPS which will work even in the presence of failure.
- h) Resource Management : Managing the huge range of resources ranging from smallest sensors to largest servers becomes critical.

III. OVERVIEW OF RESEARCH WORK DONE IN CPS

As CPS needs an interdisciplinary consideration, so it faces some problems while working on integration. As stated in section III there are various research issues. In this paper, we will discuss research challenges in following areas such as CPS composition, autonomies, architectures and resource provisioning.

A. Service Composition

Service composition means composing a service from already existing service which alone can not satisfy the requirements but in combination with other services are capable to execute a complex task. CPS is dynamic, non-deterministic so service composition needs attention [12].

Many researchers have contributed in the service composition which can be summarized as follows.

Kaiyu et. al.[13] talk about mechanisms like Monolithic, Object Oriented, Application Oriented, Component Based, Service Oriented and compared the above said mechanisms based on some metric like exhaustive analysis possible, support for reuse, support for adaption at runtime, overhead etc. Danny Huges et.al.[14] have presented Composition Challenges and Approaches which are mainly composability which is the measure of the degree to which components can be assembled in various combinations and compositionality which means that the application is compositional if the behavior of application is derived from the behavior of its constituent parts. Jian Huanag et. al.[15] have presented an ontology model for physical entity and used Artificial Intelligence planning for service composition. F. Bastani et. al.[16] have presented an efficient framework for service composition which uses 2 level composition first at abstract level and second at context level. Son et. al.[17] have considered context in the service composition and has discussed service composition in web enabled building automation system. Minyoug Kim et.

al.[18] have presented a semantic framework for reconfiguration of instrumented Cyber Physical Spaces in which they have considered an event model. Sun Meng [19] has focused on compositional modeling and composition reasoning of QoS properties.

Pascal et. al.[20] have presented a group based programming abstraction for CPS which helps to group various sensors and actuators. It facilitates feedback control mechanisms for dynamic group membership update and requirements based on feedback from the current mechanisms. It has also compared other grouping abstractions like Hood, Abstract region, Logical Neighbors, Scope etc. Swaroop et. al.[21] have presented dynamic service composition based on graph theory and service repository so dynamically a complex service can be delivered. Weining Liu [22] have presented resource aggregation and dynamic service composition framework in manufacturing domain based on runtime adjustment strategies. Martin Franke et. al.[23] have presented an approach where seamless integration of devices is possible. Peng et. al.[24] have carried out composition analysis of components and also formal verification based on service oriented architecture is discussed. Service composition also has to consider the scalability issues as there is a need to check that, considering the context of CPS whether service can be extended or not[25]. Abdul[54] has discussed cross layer automation and management model towards the dynamic composition of services in CPSs. I-Ling Yen[55] have presented novel models for specification of Physical services using Ontology for Physical Entity.Tao Wang[56] has discussed about context sensitive service composition framework with ontology model and Particle Swarm optimization technique is also proposed.

Challenges in Service Composition

The challenges in service composition can be listed as follows. Most of the papers have focused on Service Oriented Architecture which works well for Web service but in the context of CPS it needs to consider resources simultaneously. Also the methods provides a particular domain specific service composition whereas there is lack of generalized solution which can work ubiquitously. One very important point is service alone should not be considered, but has to be considered together with service, resource, context of the cyber and physical environment and dependency between resources and service. A centralized approach for service composition which faces problem from scalability point of view. And lastly the degree of compositionality and composability should be measurable in order to have a seamless service composition

B. Architectures

It is known that well-designed abstractions and architectures are critical for the success of technology[26]. Architecture is the basis in CPS design and deployment also. The different architectures proposed by various researchers can be summarized as follows.

Q Uiang Li et. al.[27] have proposed architecture based on REST style in which they have built a prototyping system called the smart gateway which integrates conceptual and physical resources into web. Jainpal Singh et. al.[28] have presented a 5-layered architecture based on Web of Things(WoT) which divides the deployment as CPS Fabric and CPS Node and mainly concentrates on event handling . Chengyuan Yu et. al.[29] have presented a 3-layered architecture based on Service-Oriented Architecture(SOA) which mainly concentrates on application re-building framework and lease protocol to guarantee atomicity. Jaipal et.al.[30] have presented 5- layered

architecture based on Web of Things with case study of Smart Home and Transportation Domain. Li Yongfu et.al.[31] have presented 5 –layered architecture based on SoA with the case study in the domain of Transportation called as T-CPS. Jing Lin et. al.[32] have presented an agent based approach for resolving data heterogeneity in the domain of Water Distribution Network. Yuchen Zhang et. al.[33] have presented 3 layer architecture with the focus of scheduling. Tao Wang et.al.[34] have presented 4 layered hierarchical network architecture with the focus of service composition . Quanyan Zhu et. al.[35] have presented 6 layered architecture with the focus on security and resilience. Son Han et. al.[36] have presented 3 layer architecture which is an enhancement of Web of Things architecture with focus on context and composition. Liang Hu et. al.[37] have presented 4 layer architecture based on SoA and talks about challenges and techniques of architecture development like real time control, security assurance and integration mechanism.

Challenges in Architectures

In the research papers[27-37], the challenges can be listed as follows. Most of the papers focused on a specific application domains (for eg. Intelligent transport, Critical infrastructure monitoring like Smart Grid, Water Distribution Network, Smart Home etc.) which works well in a particular domain, but can not be readily used in other domains. Also the architectures are designed focusing on a particular aspect of CPS like event handling, service composition, resource management, validation, verification, design, modeling etc. which works well in that particular context but not suitable in other context.

C. Resource Provisioning

Resource provisioning is defined as making the resources available as and when required irrespective of where they are and to which category they belong. Besides that physical world has continuous dynamics and cyber world has a discrete dynamics which are of different nature but still while provisioning the resources the differences needs to be resolved[38]. The CPS resource can be categorized based on various aspects. Some of them can be listed as follows.

- i) Type of a resource(Cyber, Physical or Cyber-physical) : Whether a resource is purely cyber(eg. Network Bandwidth) or purely physical (eg. Physical Sensors) or cyber-physical (eg. Data).
- ii) Mobility : Whether a resource is mobile(eg. various sensors present in mobile handset like GPS sensors, Camera Sensors deployed in Body area network etc.) or stationary sensors (which are deployed in an application specific domain eg Temperature sensor, Humidity sensors etc)
- iii) IP or non-IP enabled: Whether a resource is IP enabled full resource devices(eg. PC, laptop, Workstations, servers etc.) or IP enabled constrained resources (eg. a device with limited battery, storage , processing power, communication capability like sensors) or non -IP enabled constrained resource devices(eg. Devices like RFID tags).
- iv) Physical or virtual : Which has physical existence (eg. sensors) and virtual sensors(Virtualized sensor).

He Hua Yan et. al.[39] have presented adaptive resource management for CPS with the help of performance optimization model with resource constraint and a particle swarm algorithm is applied to solve the constraint model. A case study of unmanned vehicle with WSN navigation is presented.

Ming Li et. al.[40] have presented a scheme to collaborate with other CPS node by a cross layer optimization framework for hybrid crowd sourcing to facilitate heavy duty computation. It talks about joint computing resource management, routing and link scheduling. Kartik Lakshman et. al.[41] have presented ductility matrix to capture the mixed criticality property and presented it as a ductility maximization packing problem. Mats et. al.[42] have presented the requirement modeling at physical side. Kurt et. al.[43] have presented an ontology for resource sharing where detailed aspects of resource provisioning, resource availability, resource consumption is presented. Kai Yu Wan et.al.[44] have presented resource model to include type, definition, utility, constraints and resource mapping. They have also talked about Resource Description Template. Osman et. al.[45] have presented the scheme for categorization of optimum inter link allocation strategy that considers random attack. It is compared with regular allocation strategy. Shao et. al.[46] have presented radio resource management scheme using Cognitive radio a dynamic spectrum arrangement with the help of compressive sensing. Hai Zhuge et. al.[47] have presented a resource space model for modeling the resources .

Challenges in Resource Provisioning

Challenges faced in Resource Provisioning are as follows. Firstly the modeling method focuses only on resources and not services. Secondly no definite methods are present to check the degree of composability of the resources. Also no algorithms are provided for resource provisioning. Most of the researchers have focused on cyber resources and physical resources remain ignored. And lastly domain specific work is found.

D Autonomics

Autonomics is an inevitable characteristics of CPS. Few researchers have considered autonomies in the context of CPS but the work is in its early stage which further needs an attention of researchers.

IBM has defined the autonomic computing which has various aspects as follows[48].

- i) Self-configuration : Should automatically configure devices and network parameters as per the need.
- ii) Self-healing : Should detect, diagnose and repair localized problems
- iii) Self-optimization : Should continually seek ways to improve their operation in terms of resource utilization and to minimize various costs involved as communication cost, computation cost .
- iv) Self-Protecting : Should protect against attack which can be in the form of controlling physical device partially or completely, manipulating data which is residing in servers, databases or data in transit.
- v) Self-adaption : Should have self-adapting decisions making capability.
- vi) Self-organization : Should have self-organizing protocols (eg which are lightweight and can have group communication), should do automatic service discovery, Routing protocols which exchange data seamlessly (since CPS is inherently heterogeneous) , should maximize lifespan, should improve efficiency.
- vii) Self-Description : Self-description about characteristics & capabilities or resources
- viii) Self-Discovery : Should do dynamic discovery of service and resources
- ix) Self-energy supplying : Should have some energy conservation , energy harvesting techniques.

Ingeol et. al.[48] have introduced conceptual work on aspects of achieving autonomies in CPS. It also provides some methods for self-healing approach. Ilsun et.al.[49] have presented a framework for autonomic

computing working on MAPE (Monitor, Analyze, Plan, Execute) scheme. But the implementation details are not mentioned. It mainly concentrates self-adoption, self-healing aspect of autonomies. Levent et.al [50] have presented self-aware CPS in the application of smart buildings and cities with the help of MAPE protocol. Chonghyun et.al.[51] have presented a runtime evaluation framework for autonomies systems for self-adoption in the domain of Home surveillance. Mazier et. al.[52] have presented self-protection scheme with the help of MAPE protocol with the basic security concerns. Leon et. al.[53] have presented an autonomic reliability improvement scheme (ARIS) in smart building with the feature of self-tuning, self-managing and self-configuring.

Challenges in Autonomics

Challenges in Autonomics can be listed as follows. Firstly only conceptual frameworks are presented using MAPE protocol. Secondly as not much generalized work is done so the work found is domain specific

IV. RESEARCH PROBLEM

After studying the research paper from CPS domain we have identified research challenges in each sub domain of CPS . So the thrust area in the CPS is Resource Provisioning and Service Composition. So our research problem statement is as follows[58].

Problem definition: To provision the resources for service composition in cyber physical system with autonomies of self-management.

The overreaching objectives derived from above problem definition are as follows.

Objectives:

- 1) Composing a CPS service from the set of available any CPS services .
- 2) Provisioning the required resources for composing a new CPS service.
- 3) Achieving self-management to sustain in any critical condition and hence improving reliability.

Our future work consists of developing a middleware platform where all existing resources of various CPS will be visible and then they may participate in service composition depending upon the dependency between resource and service. Resource provisioning algorithms will run in the middleware. For middleware design, proposed approach will be multi agent based as multi agent approach[57] itself has advantages of distributedness and reuse.

V. CONCLUSION

In this paper we have given a short overview of various research challenges in the field of CPS. Among those few are discussed in detail, such as Service Composition, Resource Provisioning, and Autonomics . As no concrete work of implementation along with results is discussed in most of the papers so here we have just discussed the different approaches which researchers have considered to tackle these different problem so it does not include the comparative study of above said approaches because the case studies considered in the above papers are also different. So our scope is limited to developing generalized algorithms for Service composition and Resource Provisioning only which are not domain specific. We have developed algorithms for Service Composition and

Resource Provisioning and to validate these algorithms we have also simulated Cyber Physical Systems where runtime service is composed and as and when the resources are required , they are provisioned. The algorithms along with prototype and results will be presented as our next work.

ACKNOWLEDGMENT

I thank my supervisor Dr. Rajesh Ingle for his continuous support and encouragement throughout this research work. I also thank Dr. D. Y. Patil Institute of Technology, Pimpri, India for facilitating this research work.

REFERENCES

- [1] Eric Ke Wang,Yunming Ye, Xiaofei Xu, S.M.Yiu, L.C.K.Hui, K.P.Chow, "Security Issues and Challenges for CPS", IEEE/ACM International Conference on Green Computing and Communications & on Cyber, Physical and Social Computing, 2010.
- [2] Acatech(Ed), "Cyber Physical Systems- Driving Force For Innovation In Mobility, Energy, and Production", Acatech Position Paper, 2012.
- [3] K.-D. Kim and P.R. Kumar, "Cyber-Physical Systems: A Perspective at the Centennial," Proceedings of the IEEE, Vol. 100, Special Centennial Issue, Pages. 1287–1308, 2012
- [4] R. Poovendran, "Cyber-Physical Systems: Close Encounters Between Two Parallel Worlds," Proceedings of the IEEE, Vol. 98, No. 8, Pages. 1363–1366, 2010.
- [5] Jaipal Singh, Omar Hussain, "Cyber Physical Systems as an enabler for Next Generation Applications", Fifteenth International Conference as an Enabler for Next Generation Computing, 2012.
- [6] Borzoo Bonakdarpour, "Challenges in Transformation of Existing Real- Time Embedded Systems to Cyber Physical Systems", ACM SIGBED – Special issue on RTSS form on deeply embedded real time computing, vol 5, issue 1, 2008.
- [7] Chih Yu Lin et. al., "Enabling Cyber Physical Systems with Wireless Sensor Technologies", Article in International Journal of Distributed Sensor Networks, Vol 2012, Article id-489794, Pages 1-21.
- [8] Fang Jing Hu et. al., "From Wireless Sensor Networks Towards Cyber Physical Systems", Journal on Pervasive and Mobile Computing, Vol 7, issue 4, 2011.
- [9] Edward Lee, "Cyber Physical Systems: Design challenges", ISORC, Pages 363-369, 2008.
- [10] Raghunathan Raj et.al., "Cyber Physical Systems: The next computing revolution", International Conference on Design, Automation, 2010.
- [11] Kyoung-Dae Kim and P.R. Kumar, "An Overview and Some Challenges in CPS", Journal of the Indian Institute of Science, Vol 93, Issue3, 2013.
- [12] Edward Lee, Sanjit Sesia, "Introduction To embedded Systems- A Cyber Physical Systems Approach", UC Berkley, First Edition.
- [13] Kaiyu Wan et. al., "Investigation on Composition Mechanism for CPS", International Journal of Design, Analysis and Tools for Circuits and Systems, Vol 2, No 1, August 2011.
- [14] Danny Hughes et. al., "Composition Challenges and Approaches for CPS", International Journal of Design, Analysis and Tools for Circuits and Systems, Vol 1, No 2, 2011.
- [15] Jian Huang et. al, "A Framework for Efficient Service Composition in CPS", Fifth International Symposium on Service Oriented System Engineering, 2010 .
- [16] F. Bastani et. al., "An efficient framework for service composition for CPS", International Conference on Information Technology, Computer Engineering and Management Sciences, 2011.
- [17] Son Han et.al., "Context Aware Service Composition in Web Enabled Building Automation System", Sixth International Conference on Intelligence in Next Generation Networks(ICIN), 2012.
- [18] Minyoung Kim et.al., "A semantic Framework For Reconfiguration of Instrumented Cyber Physical Spaces", Seventh International Conference on Semantics Knowledge and Grid, 2011.
- [19] Sun Meng, "Challenges on Co-ordination For CPS", Proceedings of second International Symposium on Computer, Communication, Control and Automation(ISCCA), 2013.
- [20] Pascal et. al., "Bundle : A Group Based Programming Abstraction for CPS", IEEE transactions on Industrial Informatics, Vol 8, No 2, May 2012.
- [21] Swaroop Kalaspur et. al., "Dynamic Service Composition in Pervasive Computing", IEEE Transaction on Parallel and Distributed Systems", Vol 18, No 7, July 2007.
- [22] Weining Liu et. al., "A Solution of Dynamic manufacturing Resource Aggregation in CPS", Sixth IEEE joint International Conference on Information technology and Artificial Intelligence, 2011.
- [23] Martin Franke et. al., "A Seamless Integration ,Semantic Middleware for CPS", Tenth IEEE International Conference on Networking, Sensing and Control, 2013.
- [24] Peng Wang et. al., "Cyber Physical Systems Components Composition Analysis & Formal Verification Based on Service Oriented Architecture", Ninth IEEE international Conference on e-Business Engineering, 2012.
- [25] Kumar Padmanabh, "On the Scalability of Cyber Physical Systems", Journal of IISC, Vol 93, Issue 3, 2013.
- [26] S. Graham, G. Baliga, and P. Kumar, "Abstractions, architecture, mechanisms, and a middleware for networked control", IEEE Transactions on Automatic Control, Vol. 54, No. 7, pp. 1490–1503, 2009.
- [27] Q Uiang Li et. al., "A Case Study on REST – Style Architecture for CPS: Restful Smart Gateway", International Conference on E-Business and Information System Security, 2009.
- [28] Jainpal Singh et. al., "Event Handling for Distributed Real Time CPS", Fifteenth International Symposium on Object /component/ Service oriented Real time distributed computing workshop , 2012.
- [29] Chengyuan Yu et. al., "An Architecture of cyber physical system based on service", International Conference on computer Science and service system, 2012.

- [30] Jaipal et. al., "CPS as an enabler for next generation applications", Fifteenth international Conference on Network based Information Systems, 2012.
- [31] Li Yongfu et. al., "A service oriented architecture for the transportation cps", Proceedings of the 31st Chinese control conference, 2012.
- [32] Jing Lin et. al., "An agent based approach to reconciling data heterogeneity in CPS", IEEE international Parallel & Distributed Symposium, 2011.
- [33] Yuchen Zhang et. al., "An architecture and real time characteristics analysis of cps", Third IEEE International Conference on Communication Software and Networks, 2011.
- [34] Tao Wang et.al., "An efficient context sensitive service composition framework for precise controlling in CPS", International Conference of IT, Computer Engineering and Management Sciences, 2011.
- [35] Quanyan Zhu et. al., "Hierarchical Security Architecture for CPS", International Conference on Multimedia Technology, 2011
- [36] Son Han et. al., "Context aware service composition framework in web enabled building automation system", Sixteenth International Conference on Intelligence in Next Gen Networks, 2012.
- [37] Liang Hu et. al., "Review of CPS architecture", Fifteenth International Symposium on Object /component/ Service oriented Real time distributed computing workshop, 2012.
- [38] Paul Bogdan et. al., "Towards a Science of Cyber Physical Systems Design", IEEE/ACM Second International Conference on Cyber Physical Systems", 2011.
- [39] He Hua Yan et. al., "Adaptive Resource Management for CPS", Applied Mechanics and Materials, Vol-157-158, Pages 747-751, 2012.
- [40] Ming et. al., "Crowd sourcing in CPS stochastic optimization with Strong Stability", Forthcoming Issue In the Journal, 2013.
- [41] Kartik Lakshman et. al., "Resource Allocation in Mixed Criticality Distributed CPS", International Conference on Distributed Computing Systems, 2010.
- [42] Mats et . al., "Modeling requirements on physical side of CPS", Second International Workshop on Twin Peaks of Requirements and Architecture, 2013.
- [43] Kurt et. al., "An ontology for Resource sharing", Fifth international conference on semantic computing", 2011.
- [44] Kai Yu et.al., "Resource Modeling for CPS", International Conference on Systems and Informatics, 2012.
- [45] Osman et. al., "Optimal Allocation of Interconnecting links in CPS: Interdependence, cascading failures and robustness", IEEE transactions on parallel and Distributed Systems, Vol- 23, No 9, 2012.
- [46] Shao et. al., "Radio Resource Management for QOs guarantees in CPS", IEEE transactions on parallel and Distributed Systems, Vol- 23, No 9, 2012.
- [47] Hai Zhuge, "Probabilistic Resource Space Model for managing resources in Cyber physical society", IEEE transactions on Services Computing, Vol- 5, No 3, 2012.
- [48] Ingeol et.al., "Autonomic Computing Technologies for CPS", Eleventh International Conference on Advanced Communication Technology, 2010.
- [49] Ilsun et. al., "Autonomic computing Framework for CPS", In proc of Conference on Advances in computing, control and telecommunication technologies, 2011.
- [50] Levent et. al., "Self aware CPS in Smart building cities", Design, Automation & Test in Europe Conference & Exhibition, 2013.
- [51] Chonghyun et. al., "A Runtime Evaluation Methodology and framework for autonomies systems", ACEEE international journal on Network Security, Vol- 23, No 2, 2012.
- [52] Maziet et. al., "Towards Self metering smart metering for MAPE loop", IEEE international Symposium on a World of Wireless, Mobile and Multimedia networks, 2011.
- [53] Leon et.. al., "An autonomic reliability improvement for CPS", Fourteenth International Symposium on High Assurance Systems Engineering, 2012.
- [54] Abdul- Wahid et.al. , "Markov Task Network: A framework for Service composition under Uncertainty in Cyber Physical Systems", Journal of Sensors, Volume 16, 2016, Page 1542.
- [55] I-Ling Yen et. al. , "Rapid Service Composition Reasoning for Agile Cyber Physical Systems" , IEEE symposium on Service Oriented System Engineering, 2016
- [56] Tao Wang et.al., " Two phase Context Sensitive Service Composition Method with a Workflow model in Cyber Physical Systems" , 17th IEEE International Conference on Computational Science and Engineering, 2015.
- [57] Jing Liny et. al., "An Agent Based Approach to Reconciling Data heterogeneity in CPS", IEEE International Parallel & Distributed Processing Symposium, 2011.
- [58] Swati Nikam, Dr.Rajesh Ingle, "Resource Provisioning Algorithms for Service Composition in Cyber Physical Systems", in Proceedings of IEEE conference on Advances in Computing, Communications and Informatics (ICACCI), Pages 2797-2802, 2014

Detecting Money Laundering in a Financial System

Based on Genetic Algorithm

Ramadan Mahmood Ramo
Department of Management
Information systems
University of Mosul
rmrramo@yahoo.com

Prof. Dr. Khalil Ibrahim Alsaif
Department of Computer
University of Mosul
Sciencekhalil_alsaf@hotmai.com

Abstract

Money laundering is the process of transferring huge amounts of money gained from any of the illegal activity, such as terrorist, black money gained from non-payment of taxes to the Govt., drug trafficking etc. The money transferring process of make this illegal money gained to legal is called as "launders". The government and the financial institutions have a regulatory requirement to monitor account activity for anti-money laundering (AML) with its customers. It is therefore a mandatory requirement of any financial institution to monitor and report each money laundering activities happened under its system. The main challenge with AML is that it to track and monitor the activity of a single person, business, account, or a transaction. Therefore money laundering detection requires behavioral pattern analysis of transactions occurring over time and involving a set of related real-world entities. In this paper a Genetic Algorithm adopted in order to detect the money laundering process activities in a financial institution. The algorithm results and performance are tested in different performance metrics. The proposed approach can be fit to be used as computerized technique for money laundering.

Keyword: money laundering, genetic algorithm, clustering.

Introduction

Placement and routing are two search intensive tasks. Even though agent objects use knowledge to reduce search time, a great deal of searching is still necessary. A good proportion of this search time will be spent on optimizing the components' placement in the layout. In searching for optimum solutions, optimization techniques are used and can be

divided into three broad classes, as shown in Figure(1) [5].

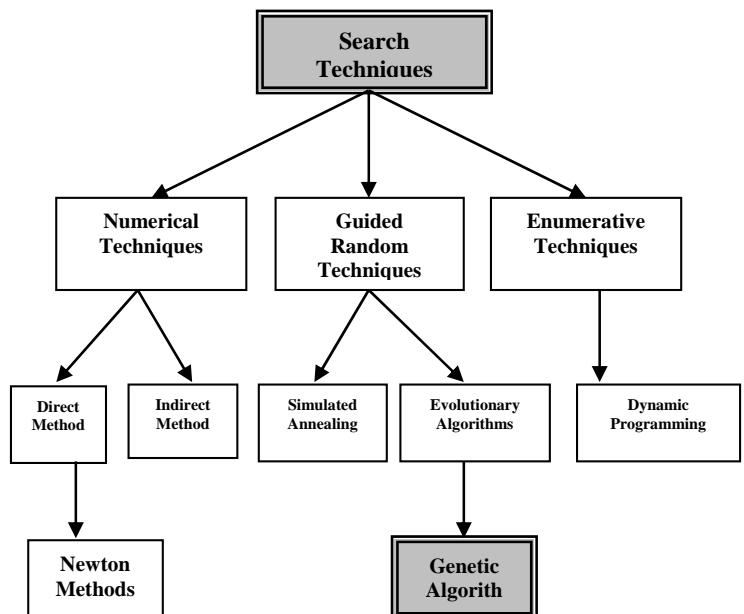


Figure (1): Optimization Techniques

Genetic Algorithms (GA) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. As such they represent an intelligent exploitation of a random search used to solve optimization problems. Although randomised, GAs are by no means random, instead they exploit historical information to direct the search into the region of better performance within the search space. The basic techniques of the GAs are designed to simulate processes in natural systems necessary for evolution, specially those follow the principles first laid down by Charles Darwin of "survival of the fittest.". Since in nature, competition among individuals for scanty resources results in the fittest individuals dominating over the weaker ones.

A financial system (within the scope of finance) is a system that allows the exchange of funds between lenders, investors, and borrowers. Financial systems

operate at national, global, and firm-specific levels. They consist of complex, closely related services, markets, and institutions intended to provide an efficient and regular linkage between investors and depositors. Money, credit, and finance are used as media of exchange in financial systems. They serve as a medium of known value for which goods and services can be exchanged as an alternative to bartering. A modern financial system may include banks (operated by the government or private sector), financial markets, financial instruments, and financial services. Financial systems allow funds to be allocated, invested, or moved between economic sectors. They enable individuals and companies to share the associated risks.[13][6]

I What Is Money Laundering.

The idea of money laundering is simple in principle. The person who has received some form of ill-gotten gains will seek to ensure that they can use these funds without people realising that they are the result of inappropriate behaviour. To do this they will need to disguise the proceeds such that the original source of the proceeds is hidden and therefore the funds themselves appear to be legitimate. Given that it is often cash that needs to be disguised, the criminal will often seek out legitimate cash-based businesses to enable them to disguise the source of their illegitimate cash. When you are discussing the laundering of money, there are generally two different connotations to consider. Money laundering refers both to the use of a cash business such as a launderette to facilitate the mingling of legal and illegal funds and also to the generic process of disguising the original proceeds of the funds, a process more normally referred to as layering. By mixing legitimate and illegitimate funds, the entire amount could potentially appear to be legitimate, and would therefore have been laundered, achieving the objectives of the money launderer. The funds will appear to have come from the legitimate business whereas some of the funds actually have arisen from criminal activity of some type. Indeed, coin-operated launderettes, which are generally cash-based businesses, would represent an ideal opportunity to achieve this, and much early money laundering did make use of legitimate cash-based activity to disguise and transform ill-gotten gains.[3][9]

II The Process Of Money Laundering

Money laundering is generally seen as a three-stage process, as shown in Figure (2).



Figure (2) : Money-Laundering Cycle

1- **Placement :** This is the movement of cash from its source. On occasion the source can be easily disguised or misrepresented. This is followed by placing it into circulation through financial institutions, casinos, shops, bureau de change and other businesses, both local and abroad. The process of placement can be carried out through many processes including[4] :

- Currency Smuggling
- Bank Complicity
- Currency Exchanges
- Securities Brokers
- Blending of Funds

2- The Layering Phase

The goal in this stage is the concealment of the criminal origin of the proceeds. Therefore, money can be transferred and split frequently between bank accounts, countries, individuals and/or corporations. Money can also be withdrawn in cash and deposited into bank accounts with other banks. It is common to use bank accounts in countries with strict banking secrecy laws and to nominate offshore corporations as the bank account holders.

3- **The Integration Phase :** Integration is the third stage of the money laundering process, in which the illegal funds or assets are successfully cleansed and appeared legitimate in the financial system.[1]

III Methods of Money Laundering

When you get down to the nuts and bolts of laundering money, there are basically only three methods to move and clean dirty money.

- Using the legitimate financial system (for example, moving money through banks, MSBs,¹ and so on)
- Physically moving the money (for example, transporting bulk cash via shipments across the border)
- Physically moving goods through the trade system.

I will describe some of the various methods of money laundering. This in no way is a complete list. Money laundering is constantly evolving, and new methods and techniques are always being developed.[10]

- Bulk Cash Smuggling.
- Gold.
- Wire Transfers
- Casinos
- Black Market Peso Exchange
- Cyber Banking
- Smart Cards
- ATMs
- Prepaid Cards
- Autos
- Correspondent Banking
- Credit Cards
- Real Estate
- Digital Currencies

IV Genetic algorithms

Genetic algorithms (GAs) are efficient, adaptive and robust search and optimization processes that are usually applied in very large, complex and multimodal search spaces. GAs are loosely modelled on the principles of natural genetic systems, where the genetic information of each individual or potential solution is encoded in structures called chromosomes. They use some domain- or problem-dependent knowledge to compute the fitness function for directing the search in more promising areas. Each individual or chromosome has an associated fitness value, which indicates its degree of goodness with respect to the solution it represents. GAs search from a set of points, called a population. Various biologically inspired operators like selection, crossover and mutation are applied on the chromosomes in the population to yield potentially better solutions.[14]

V Biological Background

The science that deals with the mechanisms responsible for similarities and differences in a species is called Genetics. The word “genetics” is derived from the Greek word “genesis” meaning “to grow” or “to become”. The science of genetics helps us to differentiate between heredity and variations and seeks to account for the resemblances and differences due to the concepts of Genetic Algorithms and directly derived from natural heredity, their source and development. The concepts of Genetic Algorithms are directly derived from natural evolution. The main terminologies involved in the biological background of species are as follows :

- *The Cell:* Every animal/human cell is a complex of many “small” factories that work together. The center of all this is the cell nucleus. The genetic information is contained in the cell nucleus. Figure 3 shows anatomy of the animal cell and cell nucleus. [15]

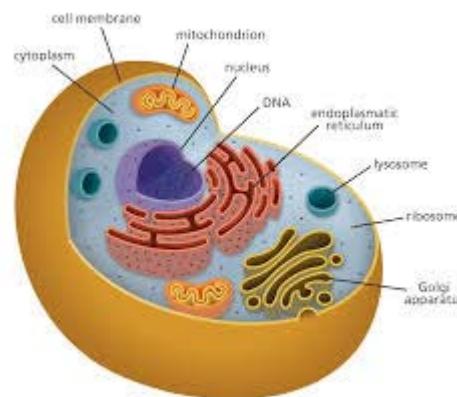


Figure (3) : the cell

- **Chromosomes:** In the nucleus of each cell, the DNA molecule is packaged into thread-like structures called chromosomes. Each chromosome is made up of DNA tightly coiled many times around proteins called histones that support its structure. Chromosomes are not visible in the cell's nucleus—not even under a microscope—when the cell is not dividing. However, the DNA that makes up chromosomes becomes more tightly packed during cell division and is then visible under a microscope. Most of what researchers know about

chromosomes was learned by observing chromosomes during cell division. Each chromosome has a constriction point called the centromere, which divides the chromosome into two sections, or "arms." The short arm of the chromosome is labeled the "p arm." The long arm of the chromosome is labeled the "q arm." The location of the centromere on each chromosome gives the chromosome its characteristic shape, and can be used to help describe the location of specific genes[16]. figure (4) represent the Chromosome.

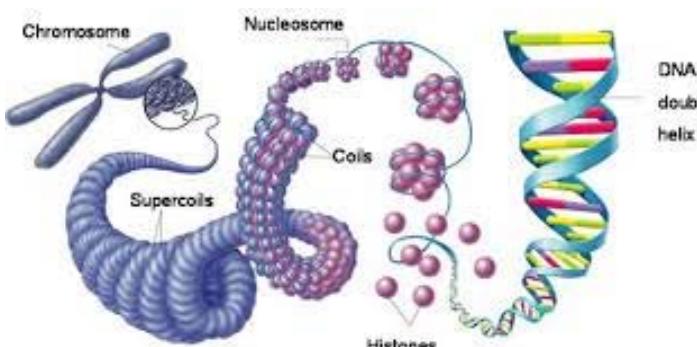


figure (4) : The Chromosome

VI Outline of Basic Genetic Algorithm[7]

1. [Start] Generate random population of n chromosomes (suitable solutions for the problem).
2. [Fitness] Evaluate the fitness $f(x)$ of each chromosome x in the population.
3. [New population] Create a new population by repeating the following steps until a new population is complete
 - ❖ [Selection] Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)
 - ❖ [Crossover] With a crossover probability cross over the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents.
 - ❖ [Mutation] With a mutation probability mutate new offspring at each locus (position in chromosome).
 - ❖ [Accepting] Place new offspring in a new population
4. [Replace] Use new generated population for a further run of algorithm

5. [Test] if the end condition is satisfied, stops, and returns the best solution in current population
6. [Loop] Go to step 2.

VII Genetic Algorithm flowchart

Above steps are visualized in the following flowchart, Figure (5).[8]

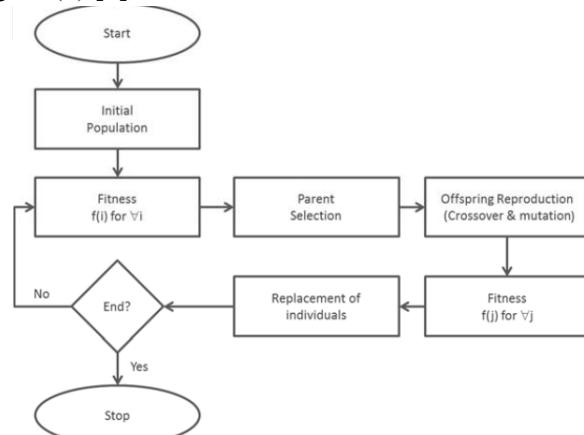


Figure (5). GA flowchart

VIII Most Of The Terms Used In Genetic Algorithm

The Table (1) show the nomenclature used in GAs is similar to the names used in biology.[5]

genotype	The code, devised to represent the parameters of the problem in the form of a string.
Chromosome	One encoded string of parameters (binary, Gray, floating point number, etc...).
Individual	One of more chromosomes with an associated fitness value.
Gene	The encoded version of a parameter of the problem being solved.
Allele	Value which a gene can assume (binary, integer).
Locus	The position that the gene occupies in the chromosome.
Phenotype	Problem version of the genotype (algorithm version) suited for being evaluated.

Fitness	Real value indicating the quality of an individual as a solution to the problem.
Environment	The problem. This is represented as a function indication the suitability of phenotypes.
Population	A set of individuals with their associated statistics (fitness average, Hamming distances, ...).
Selection	Policy for selecting one individual from the population (selection of the fittest,...).
Crossover	Operation that merges the genotypes of two selected parents to yield two new children.
Mutation	Operation than spontaneously changes one or more alleles of the genotype.

Table (1)

IX Genetic algorithm processes :

1- Selection : There are several methods used to select individuals for the implementation of the genetic algorithm on them and the following methods are seen to be used in this phase:[2]

- Roulette-wheel selection
- Random selection
- Rank selection
- Tournament selection
- Boltzmann selection
- Stochastic universal sampling

2- Crossover : is a process where two offsprings are produced by two parent chromosomes exchanging genetic information with random probability. A better offspring is expected to be created by the application of crossover operator to the mating pool .There are quite a few ways to perform crossover. Some of them are:[11]

- Single point crossover
- Two point crossover
- Multi-point crossover (N-point crossover)
- Uniform crossover
- Three parent crossover
- Crossover with reduced surrogate
- Shuffle crossover

- Precedence preservative crossover
- Ordered crossover
- Partially matched crossover
- Probabilistic crossover

3- Mutation : process is the second operator that creates new chromosomes. In mutation, spontaneous random changes are aimed so that genes that are not present in the initial population or genes that are hard to obtain from crossover operator can be created. In other words, role of mutation is to reach genes that are not present in population with random alterations. Alternatively, in mutation, a mutation rate (mutation probability) is implemented. Just like in crossover, this probability decides what percent of the population will be mutated. However, the probability of mutation should not be as high as crossover so that the population can cover the resemblance to the parent chromosomes. A high mutation rate will also lead to high randomness which will end up with too long computational time to reach the optimum solution.[12]

X Use genetic algorithm to detect money laundering

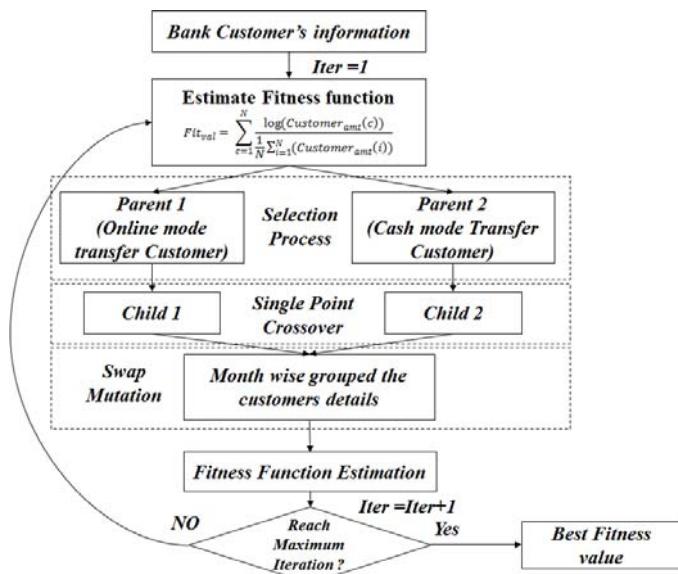
Genetic Algorithms are one of the most applied classes of algorithms for solving global/multi-modal optimization problems and have been extensively studied for solving NP-hard optimization problems. In this research the development of genetic algorithms for detecting money-laundering by

- Estimate the fitness function based on the transaction amount,
- After estimating the fitness function the parent chromosomes are generated at selection process.
- In the selection process clustered the customer by their mode of transaction.
- Next generated the child chromosomes from the above generated parent chromosome, this step is performed at the single point crossover operation.

- Then swap mutation process is carried out where the customer information is analyzed in month wise.

Again the fitness function is estimated by optimized the customer's information and found the money launders from the optimized customer's details.

XI Flowchart to detect Money laundering use GA



XII Algorithm of Detecting Money Launderin Based on Genetic Algorithm

Input: Clustered Result Cluster_{Customer}

Output: Optimized value for Detection Optimum_{value}

Procedure:

Step1: Initialize maximum iteration Max_{iter} = 10

Step2: Estimate Fitness function based on the transaction amount,

$$Fit_{val} = \sum_{c=1}^N \frac{\log(Customer_{amt}(c))}{\sum_{i=1}^N Customer_{amt}(i)}$$

Where,

Customer_{amt} –
Amount of particular clustered customer

N – Number of Transaction of the particular clustered customer

Step3: Selection process where parent chromosomes are generated,

Parent1

Parent2

Step4: Perform Single Point Cross Over,

$$CP = \text{randint}(1,1, [\text{length}(Parent1)])$$

For cc=1: CP

Child1 (cc) = Parent1 (cc);

Child2 (cc) = Parent2 (cc);

End

For cc=1: CP

Child1 (cc) = Parent2 (cc);

Child2 (cc) = Parent1 (cc);

End

// Child1 & Child2 – children chromosome

Step5: Perform swap mutation operation based on the transaction data in month wise,

$$\text{Month}_{num} = \text{Child}(\text{Date}_{trans})$$

Step6: Go to step2 for calculating the fitness value in monthly wise,

XIII The results

The output of the same ([output Collected from the source code of the Appendix](#)) is shown as below, The figure below shows the listing of Initial dataset :

Figure 6: Output list of initial dataset

-The figure 7 shows the final transaction dataset:

Figure 7: Final Transaction Dataset

-Next process is to cluster the customers report. The entire transaction details of each customer have been clustered as shown in Figure 8 below:

Figure 8: Final Transaction Dataset

-Figure 9 shows the result of this selection process:

Figure 9: Customer details based on the selection process

-Figure 10 shows the details of mutation details.

Figure 10: Month wise transaction details.

- Results of implementation of the proposed algorithm

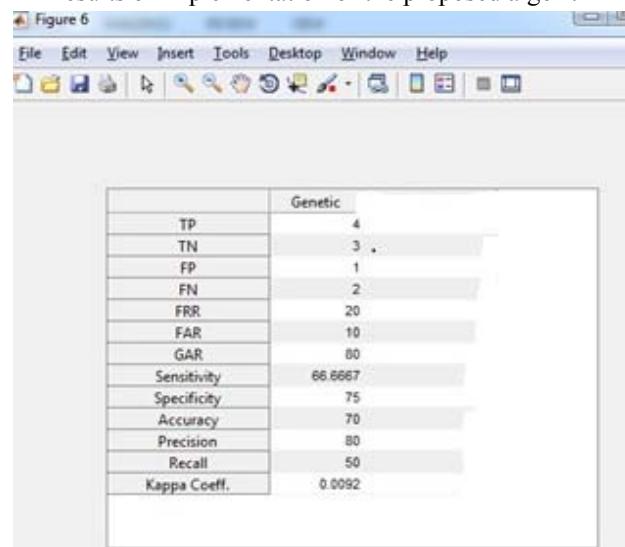


Figure 11: Performance GA

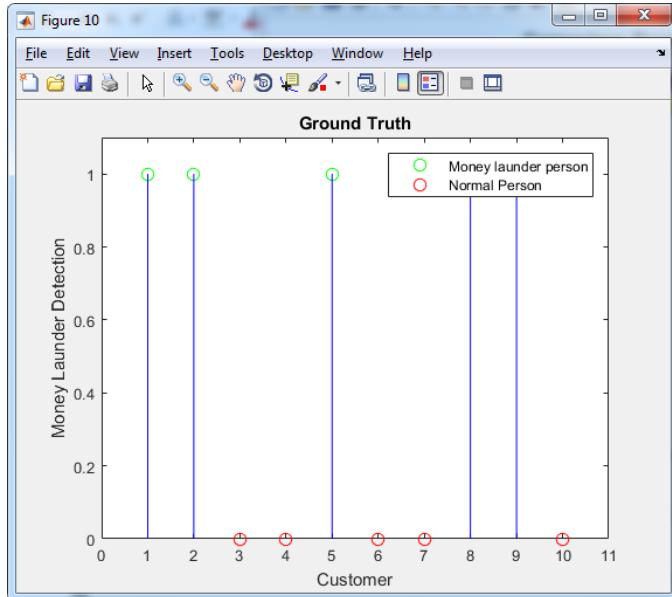


Figure 11: Comparison by ground truth

CONCLUSION

Genetic algorithms offer an excellent way of detecting and hunting money laundering activities. They should be used as described in addition to the classic rule-based detection to extend the pre-selection with two important factors. On the one hand, there won't be the problem of static detection criteria any longer. A kind of learning behavior is added to the main issue and allows quick adaptation to new ways of hiding gains from illegal business. On the other hand, the transparency of the classic methods is not any more of use to the money launderers. They will have difficulties to create transactions sets that will certainly not cause a further check by evading pre-selection criteria. The additional use of genetic money laundering detection will make life for organized crime and terrorism much harder. At least, the approach needs to be tested with a prototype application and either a batch- or real-time monitoring system have to be developed. In this research, a computer system was established to detect money laundering cases based on the genetic algorithm. The program was written using the latest version of Matlab.

Reference

- 1- AKM Ahsan," Money Laundering & Terrorist Financing Risk Management Guidelines", Bangladesh Financial Intelligence Unit BANGLADESH BANK, September, 2015, p10
- 2- Başar burak özkahya," optimization of a centrifugal compressor impeller using genetic algorithm coupled with artificial neural networks", istanbul technical university, graduate school of science engineering and technology,2016.
- 3- Dennis Cox," Handbook of Anti Money Laundering", John Wiley & Sons, Ltd, 2014, p7.
- 4- Friedrich Schneider "Money Laundering and Financial Means of Organized Crime: Some Preliminary Empirical Findings", Economics of Security Working Paper 26, Berlin: Economics of Security,2010
- 5- Fawziya M. Ramo , " Civil Objects Recognition Based on Proposed Hybrid Technique ", University of Mosul, Computer & Mathematical Sciences ,2006
- 6- Gurusamy, S., " Financial Services and systems", 2nd edition, Tata McGraw-Hill Education. ISBN 0-07-15335-3,2008.p3.
- 7- huda Abdullah Al-Omari ,2004," A Genetic Chromosomes for Image Segmentation", Msc. Thesis, Department of computer science, College of computers & mathematical Sciences, university of mosul.
- 8- Ihsan topalli," modelling user habits and providing recommendations based on hybrid television standards using artificial neural networks together with genetic algorithms", dokuz eylül university, graduate school of natural and applied sciences,2017,p16
- 9- John madinger," money Laundering A guide for criminal investigators", taylor & francis group, llc,2012,p5
- 10- Kevin Sullivan," Anti-Money Laundering in a Nutshell", Awareness and Compliance for Financial Personnel and Business,2015, p14
- 11- Maulik, U., & Bandyopadhyay, S."Genetic algorithm-based clustering technique. Pattern Recognition" , 33, 1455-1465.2000.
- 12- Mustafa gürel," a hybrid genetic algorithm for multi mode resource constrained scheduling problem for large size projects",middle east technical university, the graduate school of natural and applied sciences,2015,p37
- 13- O'Sullivan, Arthur; Sheffrin, Steven M. (2003). Economics: Principles in Action. Upper Saddle River, New Jersey 07458: Pearson Prentice Hall. p. 551. ISBN 0-13-063085-3.
- 14- Sanghamitra Bandyopadhyay,Sankar K. Pal," Classification and Learning Using Genetic Algorithms", Springer-Verlag Berlin Heidelberg, 2007, p21.

15- S.N.Sivanandam · S.N.Deepa," Introduction to Genetic Algorithms", Springer-Verlag Berlin Heidelberg 2008,P16

16- Wolfram, Antonin; Neumann, Heinz "Chromosome condensation and decondensation during mitosis". Current Opinion in Cell Biology. Elsevier Ltd. ,2017.

Appendix (Source Code)

```
[parent1_type,parent2_type,child_trans,Date_trans,cus_month,Total_amt,Genetic_res]=Basic_Genetic(customer_detail)
% Evaluate Fitness Value
fitval=fitness_func(customer_detail);
% Selection Process
field='Transaction Type';
str2=strcmp(customer_detail(1,:),field);
Type_trans=customer_detail(2:end,str2);
Num_type=unique(Type_trans);
xx=2;
yy=2;
parent1_type(1,:)=customer_detail(1,:);
parent2_type(1,:)=customer_detail(1,:);
for ii=1:length(Type_trans)
    if strcmp(Num_type{1},Type_trans{ii})
        Type_trans1(ii)=1;
    else
        Type_trans1(ii)=2;
    end
end
res=find(Type_trans1==1);
for ii=1:length(res)
    parent1_type(ii+1,:)=customer_detail(res(ii)+1,:);
end
res1=find(Type_trans1==2);
for ii=1:length(res1)
    parent2_type(ii+1,:)=customer_detail(res1(ii)+1,:);
end
function
cluster_customer=cluster_method(Preprocessed_trans)
field='Customer Id';
strr=strcmp(Preprocessed_trans(1,:),field);
customer_info=cell2mat(Preprocessed_trans(2:end,strr));
Num_customer=unique(customer_info);
for ii=1:length(Num_customer)
    cust_loc=find(customer_info==Num_customer(ii));
    x=2; for kk=1:length(cust_loc)
        cluster_customer{ii}(1,:)=Preprocessed_trans(1,:);
        cluster_customer{ii}(x,:)=Preprocessed_trans(cust_loc(kk)+1,:);
    end
end
```

```
x=x+1;
end
end
function [Total_amt,cus_amt]=fitness_func(custo_cal)
field='Transaction Amount';
str1=strcmp(custo_cal(1,:),field);
customer_amt=cell2mat(custo_cal(2:end,str1));
Total_amt=0;
for ii=1:length(customer_amt)
    Total_amt=Total_amt+customer_amt(ii);

    cus_amt(ii)=log(customer_amt(ii))./mean(customer_amt);
end
cus_amt1=sum(cus_amt);

% Single Point Cross Over
siz=round(size(parent1_type,1))-1;
siz1=round(size(parent2_type,1))-1;
child1=parent1_type(2:siz,:);
child2=parent1_type(siz+1:end,:);
child11=parent2_type(2:siz1,:);
child21=parent2_type(siz1+1:end,:);
child_1=[child1;child11];
child_2=[child2;child21];
child=[child_1;child_2];
size_chi=size(child,1);
child_trans=customer_detail(1,:);
child_trans(2:size_chi+1,:)=child;
% % Swap Mutation
field='Transaction Date';
str3=strcmp(customer_detail(1,:),field);
Date_trans=child_trans(2:end,str3);
for ii=1:length(Date_trans)
    datet=Date_trans{ii};
    strs=strsplit(datet,'/');
    month_trans(ii)=str2num(strs{2});
end
%
uni_mon=unique(month_trans);
for ii=1:length(uni_mon)
    mon_par=find(uni_mon(ii)==month_trans);
    cus_month{ii}(1,:)=child_trans(1,:);
    for kk=1:length(mon_par)
        cus_month{ii}(kk+1,:)=child(mon_par(kk),:);
    end end % Fitness Function call
for ii=1:length(cus_month)
    cust_mon_trans=cus_month{ii};
    [Total_amt,cus_amt]=fitness_func(cust_mon_trans);
    Genetic_res(ii,:)=cus_amt;
end
```

Design of a New Small Antenna for Passive UHF RFID Tags

Ines Frigui, Mohamed Salah Karoui, Hamadi Ghariani, Mongi Lahiani

Laboratory of Electronics and Technologies of Information (LETI)

National School of Engineers of Sfax (ENIS)

Sfax, Tunisia

Abstract—In this paper, a small antenna for a passive radio frequency identification (RFID) transponder operating in the ultra-high frequency (UHF) band is introduced. Thanks to the combined miniaturization techniques: remodeled form of both the meandered dipole and the T-matching, a transformed ground plane and finally two shorting plates, the designed antenna on Rogers RT5880 operates at 915 MHz with an overall size of only 31mm x 22.75mm x 1.575mm and provides a good impedance matching, a wide impedance bandwidth and a positive maximum gain at the resonant frequency.

Keywords-small antenna; passive RFID transponder; UHF band; miniaturization techniques; meandered dipole; T-matching; ground plane; shorting plates.

I. INTRODUCTION

In recent years, radio frequency identification (RFID) technology has been extensively used in various applications for tracking products in transit, goods, animals and people all over the world due to its advantageous identification procedure based on radio frequency waves [1].

Distinct ultra-high frequency (UHF) band can be devoted according to the regulations in the work region. To cover the entire passive UHF band, the allocated frequency range must be between 860 to 960 MHz [1].

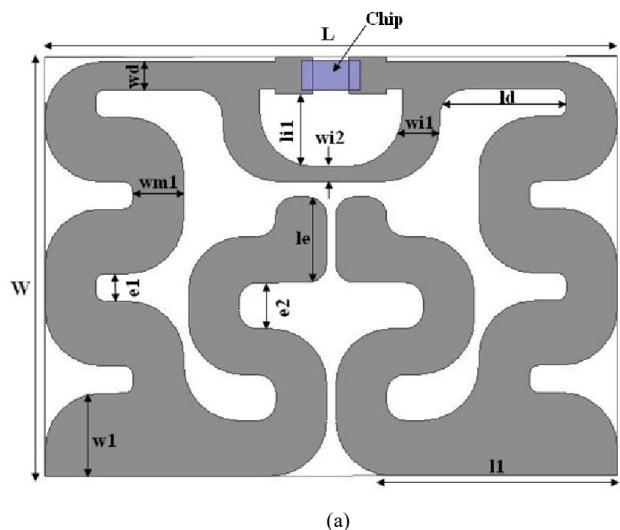
In a passive RFID system, the tag consists of an antenna integrated with a memory chip that saves information about the object to be attracted to. Actually, the large scale implementation of RFID tag highly requires that the tag must be smaller in size and easy to manufacture in order to be implemented at enormously large scale. Moreover, the difficult challenge in the tag antenna designing is to ensure the greatest size reduction without performances degradation [2], [3].

In this paper, a new compact symmetric passive UHF tag antenna operating at 915 MHz is presented. It is composed of a radiating modified meandered dipole strips, an altered T-matching and a modified ground plane. Each folded dipole arm is shorted to the suggested ground plane via shorting plate.

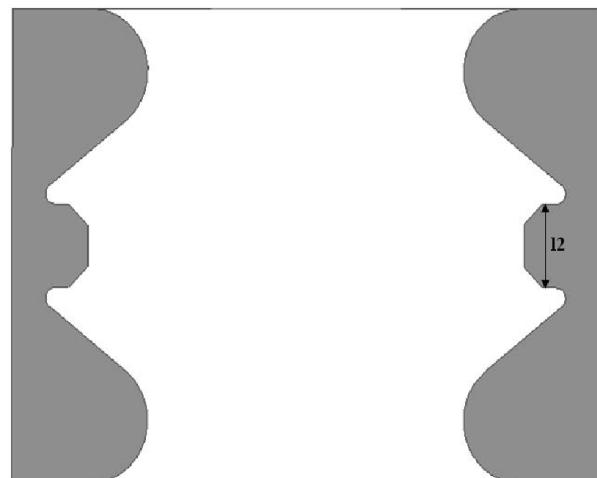
II. TAG ANTENNA CONFIGURATION

The structure of the new designed tag antenna is illustrated in Fig. 1. The proposed antenna is implemented on a double sided Rogers RT/Duroide 5880 substrate (dielectric constant $\epsilon_r=2.2$, loss tangent $\tan(\delta)=0.0009$) with a thickness of

1.575mm and etched with 0.035mm copper thick. It consists of meandered dipole arms and an inductively coupled feed line both are with fillet vertices, in the top view of the substrate. A modified ground structure, in the back side. The radiating element is short-circuited to the ground plane with two symmetric shorting plates.



(a)



(b)

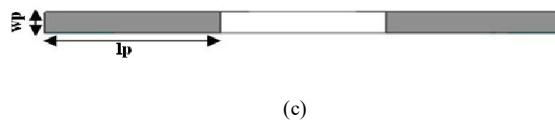


Figure 1. Geometry of the proposed UHF Tag antenna: (a)Top view, (b)Bottom view and (c)Side view

III. SIMULATION RESULTS AND ANALYSIS

A. Parametric Study

First of all, to control the influence of the different antenna geometric parameters on both the resonant frequency and its peak gain, a parametric study using the high frequency simulation software (HFSS) v.13 is done.

In this study, we just change a single parameter at one time (w_{i1} or w_{i2} or l_p or w_p) while others stay unaltered ($L=31\text{mm}$, $W=22.75\text{mm}$, $w_d=1.5\text{mm}$, $l_d=6.4357\text{mm}$, $l_{i1}=3.9625\text{mm}$, $w_{m1}=2.75\text{mm}$, $l_1=12.25\text{mm}$, $l_e=4.65\text{mm}$, $w_1=4.5\text{mm}$, $e_1=1.5\text{mm}$, $e_2=2.5\text{mm}$, $l_2=4\text{mm}$).

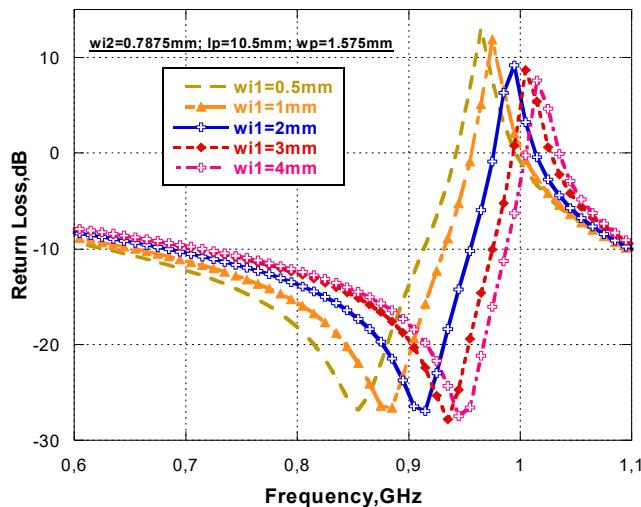


Figure 2. Variation of the return loss for different w_{i1} values.

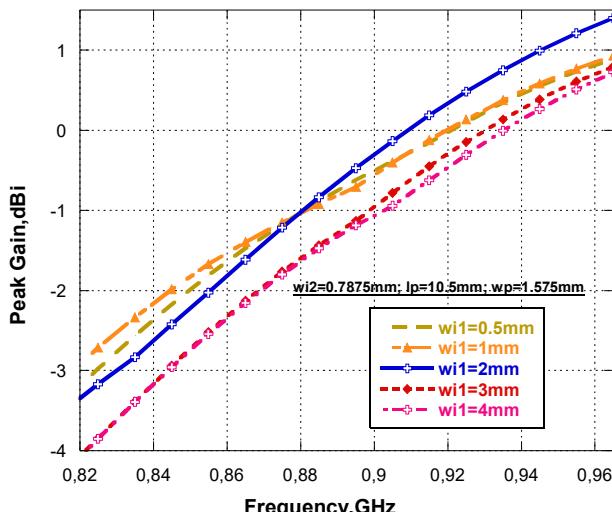


Figure 3. Variation of the peak gain for different w_{i1} values.

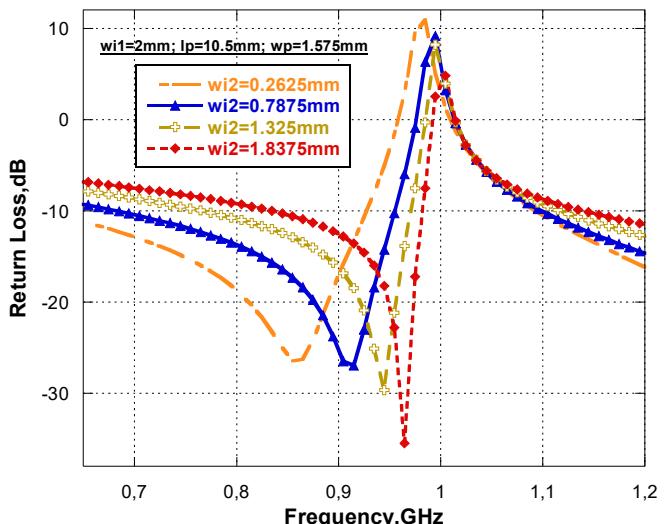


Figure 4. Variation of the return loss for different w_{i2} values.

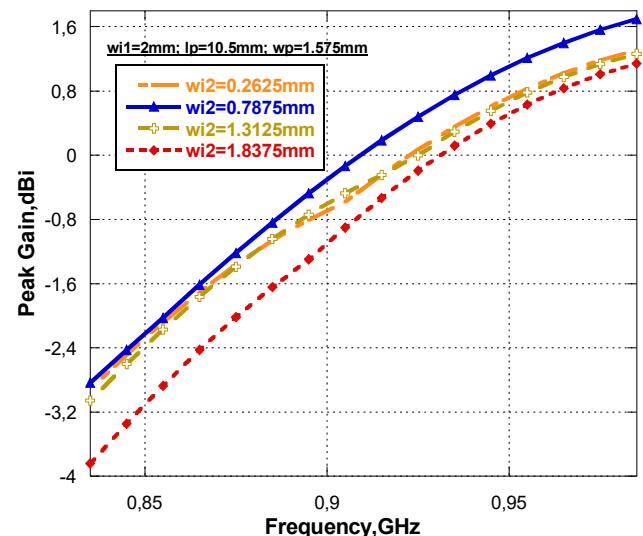


Figure 5. Variation of the peak gain for different w_{i2} values.

The dimensions of the width w_{i1} and w_{i2} of the inductive loop are of a high importance while designing the proposed antenna. As shown in Fig. 2, Fig. 3, Fig. 4 and Fig. 5, the increase in w_{i1} or w_{i2} increases both the central frequency and its peak gain.

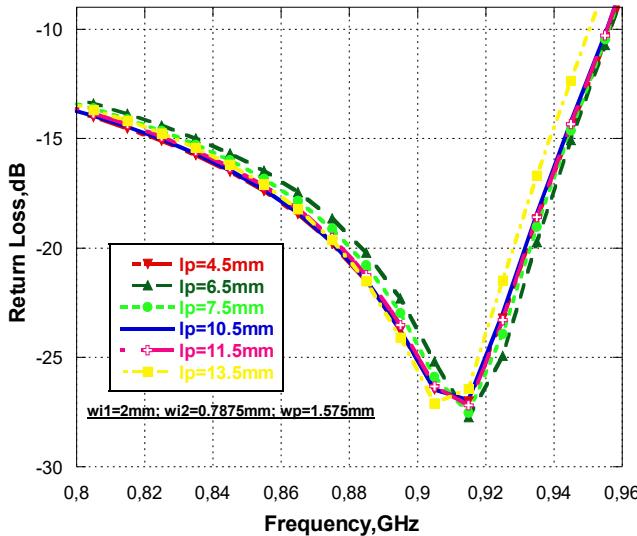


Figure 6. Variation of the return loss for different lp values.

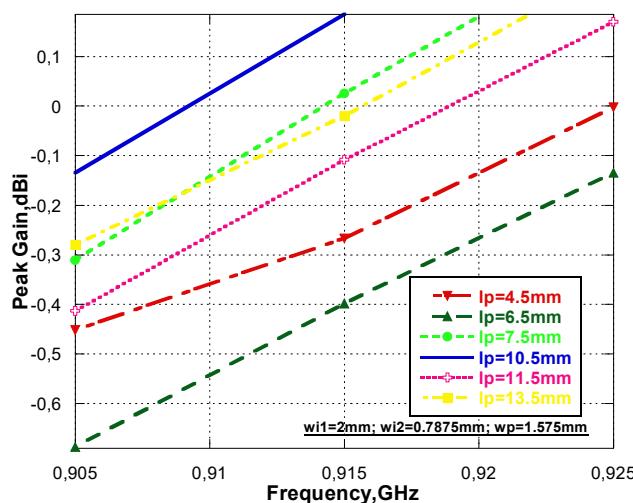


Figure 7. Variation of the peak gain for different lp values.

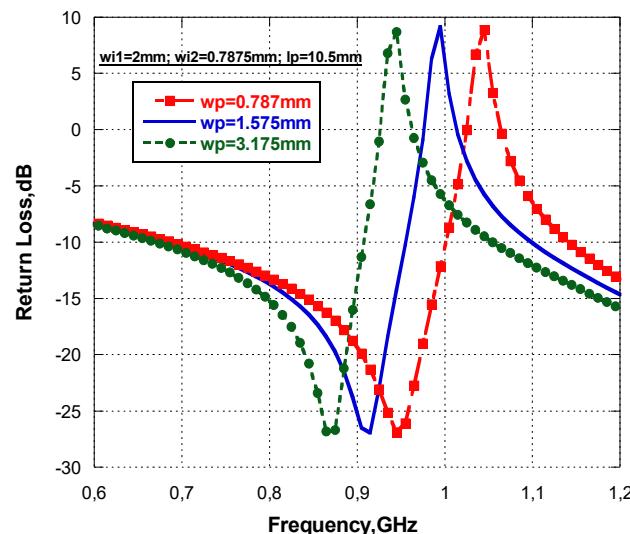


Figure 8. Variation of the return loss for different wp values.

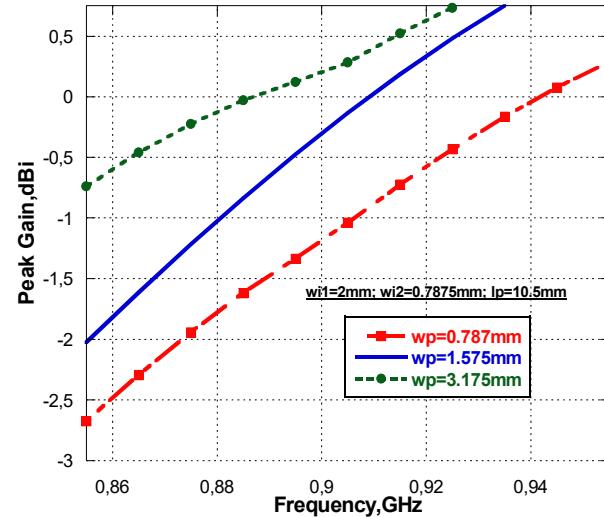


Figure 9. Variation of the peak gain for different wp values.

The shorting plate size (wp, lp) obviously effect on the novel antenna characteristics. From Fig. 6 and Fig. 7, it is noted that the shorting plate length influences the tag peak gain more than its center frequency. However, modifying the width of the rectangular shorting element, which means modifying the thickness of the substrate, shifts the tag resonance frequency and its maximum gain values at the same time as observed in Fig. 8 and Fig. 9.

B. Complementary Simulations

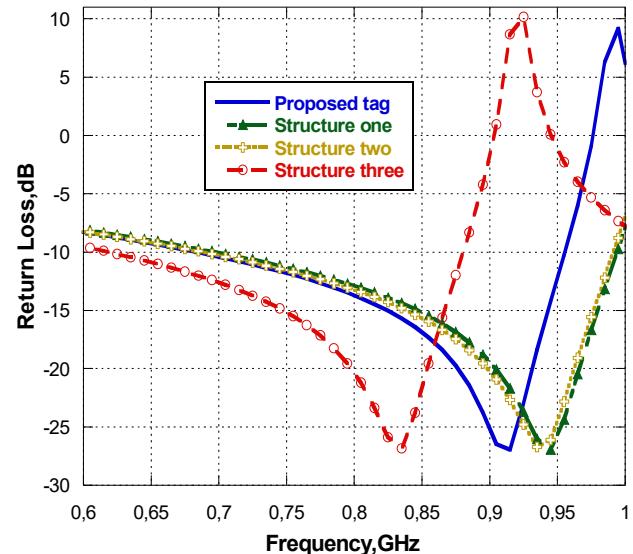


Figure 10. Return loss behavior comparison between the proposed tag and three other different structures.

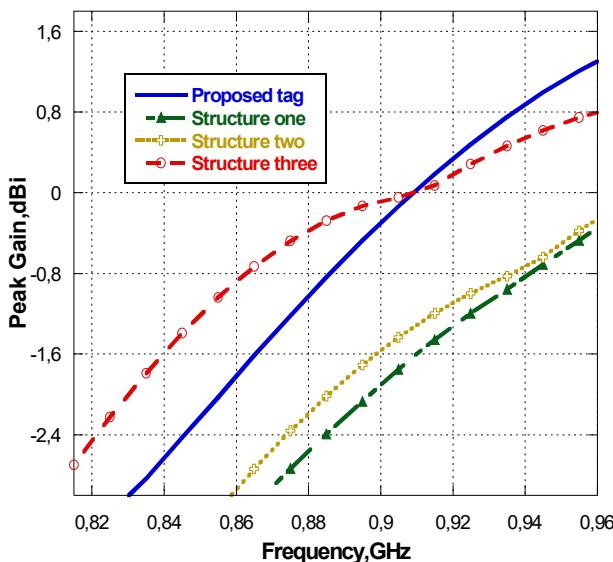


Figure 11. Peak gain behavior comparison between the proposed tag and three other different structures.

Fig. 10 and Fig. 11 present a comparison between the proposed tag and three other different structures in terms of operating frequency and its peak gain values. In fact, structure one (Fig. 12.(a)) presents the proposed tag without the shorting plates. Structure two (Fig. 12.(b)) presents the new tag without the ground plane and finally structure three (Fig. 12.(c)) presents the proposed tag without fillet vertices of the radiating element.

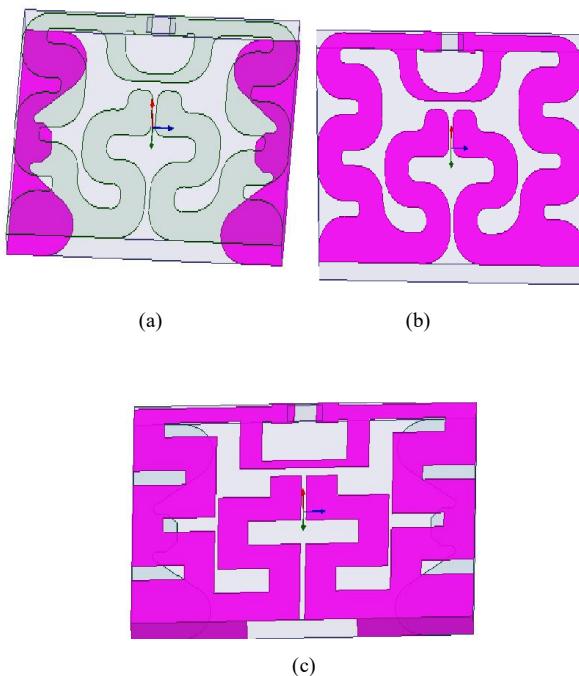


Figure 12. Three other different structures: (a) Structure one, (b) Structure two and (c) Structure three.

- The comparison between the novel UHF tag and structure one demonstrates that inserting the shorting plates permits a decrease in the center frequency from 945 MHz to 915 MHz and what is more interesting that the peak gain at the center frequency goes from -0.71 dBi to 0.18 dBi.
- The comparison between the novel UHF tag and structure two proves the efficiency of the suggested modified ground plane. Indeed, this ground plane allows a resonant frequency shift to lower values, the resonant frequency goes from 935 MHz to 915 MHz, and what is more important that it also allows a significant gain improvement of the order of 1.01 dBi.
- The comparison between the novel UHF tag and structure three confirms that the altered form (rounding the corners of rectangular form) of the meandered dipole arms and the inductive loop, is very required to get a serious peak gain enhancement [4] with about 1.97 dBi.

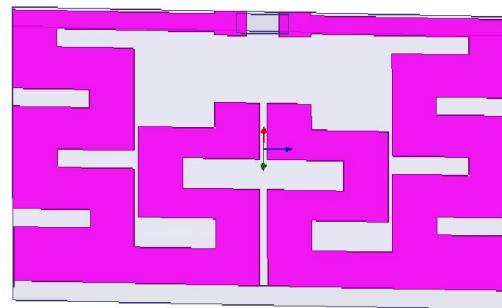


Figure 13. The conventional meandered dipole arms structure.

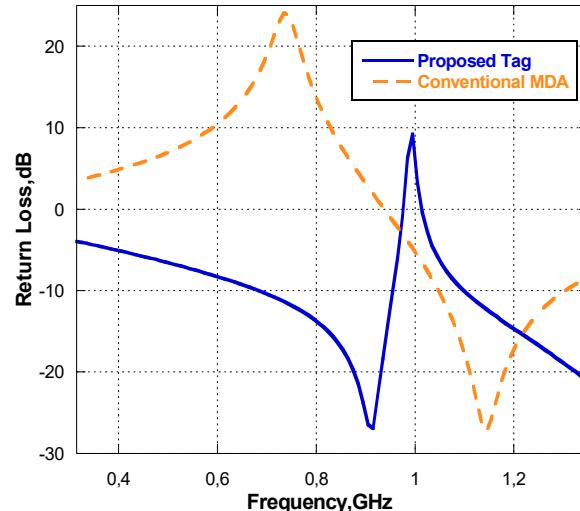


Figure 14. Return loss behavior comparison between the proposed tag and the conventional meandered dipole arms structure.

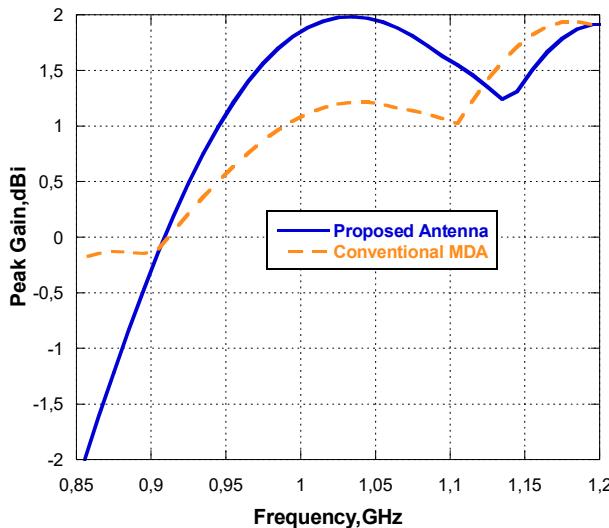


Figure 15. Peak gain behavior comparison between the proposed tag and the conventional meandered dipole arms structure.

- A final comparison between the proposed tag and a simple conventional meandered dipole arms (Fig. 13) is set. As depicted in Fig. 14 and Fig. 15, for the same size, the new designed tag antenna reduces the resonant frequency from 1145 MHz to 915 MHz and the peak gain from 1.71 dBi to only 0.18 dBi. The combined three techniques: modified ground, the inductive loop and meandered dipole arms fillet in vertices and the shorting plates strongly decrease the resonant frequency to the desired value with still an acceptable positive gain.

C. Optimized Tag Antenna Simulation Results

The antenna performances with optimal parameters cited in Table I were first evaluated by using (HFSS) v.13.

TABLE I. OPTIMIZED TAG ANTENNA GEOMETRIC PARAMETERS

Parameter	Dimension (mm)	Parameter	Dimension (mm)
<i>L</i>	31	<i>lI</i>	12.25
<i>W</i>	22.75	<i>le</i>	4.65
<i>wd</i>	1.5	<i>wI</i>	4.5
<i>ld</i>	6.4357	<i>eI</i>	1.5
<i>wi1</i>	2	<i>e2</i>	2.5
<i>wi2</i>	0.7875	<i>l2</i>	4
<i>li1</i>	3.9625	<i>wp</i>	1.575
<i>wm1</i>	2.75	<i>lp</i>	10.5

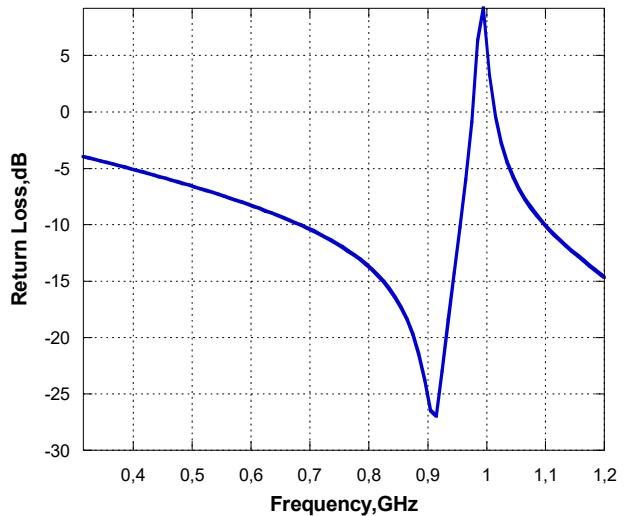


Figure 16. HFSS simulated return loss of the optimised UHF Tag.

As observed in Fig. 16, the proposal antenna operates at a resonance frequency of 915 MHz with a satisfactory return loss of -26.95 dB. The impedance bandwidth is estimated to (955-682) MHz = 273 MHz.

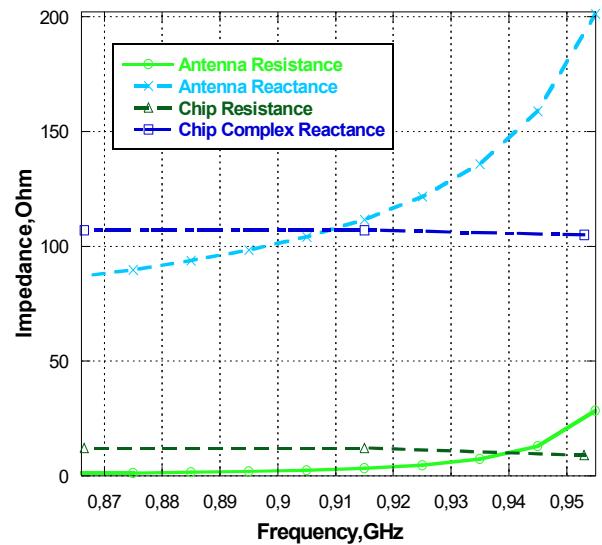


Figure 17. Input impedances in a region of interest.

Fig. 17 is another proof that the antenna has a good adaptation. In fact, the antenna input impedance is equal to $(3.26+j111.51) \Omega$ at 915 MHz which is close to the complex conjugate of the input impedance of the employed MURATA RFID Magicstrap LXMS31ACNA-010 chip [5] which has an input impedance of $(12-j107) \Omega$ at the same frequency.

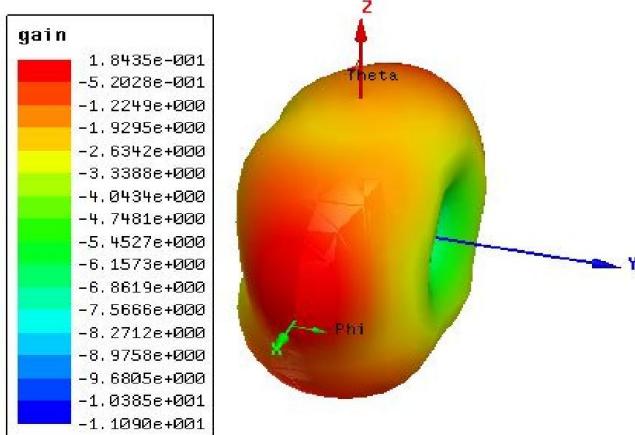


Figure 18. Maximum 3-D gain of the optimised UHF Tag at 915MHz.

Fig. 18 represents the maximum 3-D gain of the tag antenna which is about 0.18 dBi at the operating frequency.

The use of a second simulation tool is highly recommended to validate the results already been retrieved. so the antenna was also evaluated using computer science technology (CST) v.14.

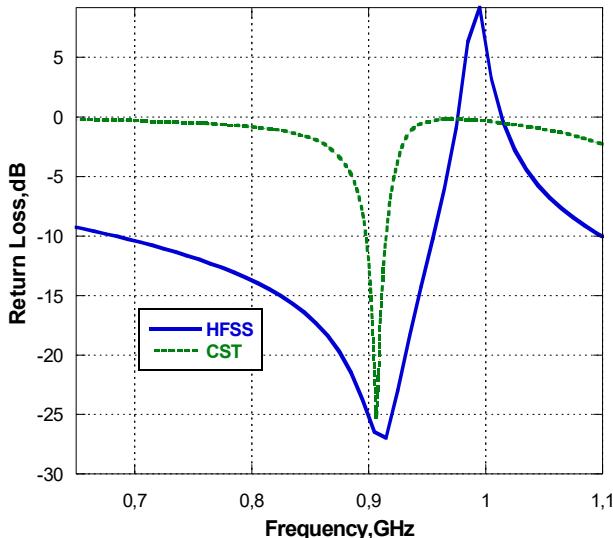


Figure 19. HFSS and CST simulated return loss comparison.

From Fig. 19, it seems that there is an agreement between the two simulation results. A slight difference is detected. It is mostly justified by the difference between the mathematical methods of each simulation tool solver. The first software uses the finite element method (FEM) whereas the second one uses the finite integral technology (FIT).

TABLE II. PROPOSED UHF TAG ANTENNA DIMENSION STATE OF THE ART

Antenna	Substrate	Resonant Frequency (MHz)	Dimension (mm ³)
[3]	Paper	915	76.2 X 25.4 X 0.25
[6]	Rogers RT5870	915	40 X 40 X 0.8
[7]	Rogers RT5880	921	58 X 34X 0.85
[8]	FR4	915	68.8 X 32.5 X 0.5
[9]	FR4	925	67 X 50 X 1.6
[10]	FR4	920	77.68 X 35.5 X 1.6
Proposed	Rogers RT5880	915	31 X 22.75 X 1.575

If we compare our work to other recent existing UHF tags Table II, we find that our designed antenna is a good candidate for miniaturized UHF tags. In fact, the motivation of this work is that it permits the smallest size with still acceptable characteristics as it was previously discussed and detailed.

IV. CONCLUSION

A novel compact antenna appropriate for RFID applications has been suggested. All requirements for passive UHF RFID transponder were satisfied with the advantages of simple structure, considerable size reduction, good adaptation, wide bandwidth and an acceptable positive gain.

REFERENCES

- [1] Nemai Chandra Karmakar, Handbook of Smart Antenna for RFID Systems (John Wiley & Sons, Inc, 2010).
- [2] Ming-Tao Zhang, Yong-Chang Jiao, Fu-Shun Zhang and Wu-Tu Wang (2009). Design of Antennas for RFID Application, Development and Implementation of RFID Technology, Cristina Turcu (Ed.), ISBN: 978-3-90261354-7, InTech, Available from: https://www.intechopen.com/books/development_and_implementation_of_rfid_technology/design_of_antennas_for_rfid_application.
- [3] Abdul Quddious, Farooq A. Tahir, Hammad M. Cheema , "A Novel Printed RFID Tag Antenna for Specific UHF Bands", Proceedings of 12th International Bhurban Conference on Applied Sciences & Technology (IBCAST), Islamabad- Pakistan, January2015.
- [4] Mohamed Salah KAROUI, Hamadi GHARIANI, Mounir SAMET, " STUDY AND DESIGN OF A LOOP ANTENNA FOR MEDICAL TELEMETRY APPLICATION", Third International Conference on Systems Signals & Devices, Sousse-Tunisia, March 2005.
- [5] Murata MAGICSTRAP® Application Note, Online Available: <http://www.murata.co.jp/products/rfid/index.html>.
- [6] S. BELDI, A. GHARSALLAH,Tan Phu Vuong , "A Miniatureized Tag Antenna RFID for Textil Application", Computer & Information Technology Global Summit (GSCIT), Sousse-Tunisia, June 2014.
- [7] N.M.Faudzi, M.T.Ali, I. Ismail, H.Jumaat, N.H.M.Sukaimi,"Compact Microstrip Patch UHF-RFID Tag Antenna For Metal Object", Symposium on Wireless Technology and Applications (ISWTA), Kota Kinabalu-Malaysia, 2014.
- [8] Yan Shi, Chao Fang, Kang Qi, and Chang-Hong Liang," A Broadband Design of UHF Fractal RFID Tag Antenna", Progress In Electromagnetics Research Letters, Vol. 58, pp. 45-51, 2016.
- [9] Pouria Kamalvand, Gaurav Kumar Pandey and Manoj Kumar Meshram," A single-sided meandered-dual-antenna structure for UHF RFID tags", International Journal of Microwave and Wireless Technologies, pp. 1-8, Cambridge University Press and the European Microwave Association, 2017.
- [10] Md. Rokunuzzaman, Mohammad Tariqul Islam, Wayne S. T. Rowe, Salehin Kibria, Mandeep Jit Singh, Norbahiah Misran," Design of a Miniaturized Meandered Line Antenna for UHF RFID Tags", PLoS ONE 11(8), August 2016.

Efficient Multi-Level Authentication for Cloud API based on RestPL

M.J.Balachandran
Research Scholar
School of Computing Sciences
Vels University
Chennai, India
mjbalachandran@yahoo.co.in

Abstract – Objective: Nowadays lot of cloud computing systems with web applications uses representational state transfer (REST) API deployed for their simplicity services. The implementation of the new REST system acts as protocol, it is explicitly for managing and exchange of data in Internet services which is completely transformed the software development after 2000. Currently all the companies or application has a REST API for business development. Today there are no projects or an application that doesn't have a REST API for the creation of professional services based on this software. Interface of a user from the server and data storage are separated by REST. REST API is a type of independent language or platform.

Methods/Statistical Analysis: This paper detailed the ways of how efficiently authenticated through Cloud API in Jelastic Server based on RestPL. RestPL is purely based on the initiation of request. **Findings:** This application work involves 6 stages, Jelastic Server module, Root Signature generation, Send Root Signature, File sharing in Jelastic, User File access in jelastic & File Sustainability.

Applications/Improvements: Boundary splitting algorithm is utilized to split the fingerprint image into eight parts and Merkle Hash Trees takes input as these splited eight parts, so that it creates root signature for the provided fingerprint image. Further research will involve iris or face recognition.

Keywords: Application programming interface, insecure API, Activity role based Access & Cloud computing Challenges.

I. INTRODUCTION

An architecture which specifies about a constraint like uniform interface during the application of a web service triggering desirable properties which includes changes, performance and scalability that enable services to efficiently works on the internet is a Representational State Transfer (REST).

Features of RESTFUL services : When a system uses resources like information of business, web pages, A/V files, images or anything that can be represented in a computer-based system. Service

Dr.P.Sujatha
Associate Professor
Department of Computer Science
Vels University
Chennai, India
suja.research@gmail.com

provider's intention is to provide a space or a window to their customers so that the accessing of resources is done by them. Technical experts are expecting these services can be easily scalable, extensible and maintainable. The design of RESTful gives assurance of the above and more. In general, services of RESTful has the properties/features like Representations, Messages, URIs, Uniform interface, Stateless, Links between resources & Caching.

Below are the scenarios chosen to expose REST API when, (i) We don't know the end user (client) stack. (ii) Interfacing with multiple systems (iii) We need the cost benefit by offloading the task of platform specific integration up to the API consumers. (iv) Not attentive whether our API is consumed technically.

(v) When we require universal presence with minimum efforts, given the fact that REST API is an HTTP Service which is virtually exists on almost all the platforms.

Controlling of access is done through enforcement of some policy to the provided interface, this is done by the service providers for the security purposes. Even though the access is controlled by the Cloud Service provider, it still lacks in policy language, efficient authentication of REST interfaces, hence it brings couple of key limitations:

First, to accommodate certain systems, consumers to deal with totally various types of policies. Secondly, provision of own design architecture is to be done by the Cloud service providers. It is a platform-specific authorization policy language and its respective enforcement mechanisms. This paper detailed the ways of how efficiently authenticated through Cloud API in Jelastic Server based on RestPL. RestPL is based on request oriented.

The threats can be launched in cloud, the modernized ways to exploit these threats and their impacts on

cloud. Based on the threats and its respective analysis presented the resolutions of security for the prevention of these threats by implementing an application. The reason is to implement a resolution to the growing trend of attacking through an API.

II. LITERATURE SURVEY

In this paper, Yang Luo, Hongbo Zhou, Qingni Shen, Anbang Ruan, Zhonghai Wu [1], proposed a REST Policy Language (RestPL) which explains on policy authorization especially for REST APIs. RestPL is request oriented and some standard request form has been defined wherein it indicates that this policy generates on its own from an actual request. The eXtensible Access Control Markup Language (XACML) is a popular standard proposed by OASIS which assures platform-independent policies for access control. Here, authors have designed an application which is based on a particular request on access control language named as REST Policy Language (Rest- PL), this has caused the reduction of load for both the Cloud service providers and consumers. Standard request form designed by the authors for REST, RestPL can be easily implemented like a cloud computing platform by a service provider. Based on the experimental results, there is a reduction of RestPL overhead from 80.6% when compared with the original policy.

Stine Lomborg, Anja Bechmann [2] elaborates how it is social media is benefitted in availing the data for the researchers through their application programming interfaces (APIs). By using an API which acts through back-end wherein the connection of new add-ons to an existing service by the third-party developers. Displaying a challenging discussion of methodology of the opportunities, research based on quantitative and qualitative on APIs, this paper highlighted methodological issues during the collection and validation of data through APIs. Here, author also explained the advantages of using the APIs and other techniques of computation prompting integration of quantitative as well as qualitative analysis during collaboration of methods on research designs.

Muhammad Imran Hussain and Naveed Dilber [3] explains about the challenges of restful web services and its respective security issues as it does not have either predefined security standards or methods nor own security models. The importance of this research is to exploring ASP.NET web APIs using MVC framework for the implementation of restful web service. Using various security techniques like token, claim based, web token, OAuth2 and delegation based since it has constraints of H2serious because it is

client-server, stateless and cashable etc. Here, author explained, how web services based on REST are communicated using MVC framework using XML, JSON and plain text as an exchange data format by using ASP.Net Web API. He has explained in detail that the practice of securing the REST API web services by using ASP.NET MVC Web API.

Jyoti Joshi [4] explained about insecure API, encryption, access control and authentication and task monitoring, the mentioned interfaces designed to secure against malicious and accidental attempts to get around the policies. In this paper author is proposing an access control mechanism using the JV-Role Based Access Control Model (JV-RBAC), The implementation of access control policy at the API level in JV-RBAC. Attributes is taken along with the identification when the user is provided access and authenticated, Cloud service or domain name can be accessed by the IP address of the machine. The proposed model enables three layers of API security system. Green zone layer allows or provide access to registered users alone, from this zone cloud service can be accessed but simultaneously required input to be obtained for stage 2 and 3 can be accessed based on the output of stage 2 which will act as an input to stage 3. Author has chosen JV-RBAC Model because it suits best for the business operations and large enterprise needs. This provides assurance or enables easy way for the firm or corporate to map a user's local organization role with the global role wherein the access of services are ensured.

Navid Pustchi, Ravi Sandhu [5], has presented an access control model based on attributes. This model ensures and enabling association among the tenants of a cloud provider and cloud systems. Here, it allows single cloud to collaborative access control models because the requirement restricts to collaboration of cross-tenant and deferring of cross-cloud integration issues. For sharing of available resources, author is suggesting an enablement of collaboration through value assignments of multiple tenants attribute supported by the cloud service provider. In that approach, necessary attributes to tenants which is a trusted one where authorization provided to the tenants of trustee to assign the values of attribute to user attributes of a trustor tenants. In that approach, post separating the attributes to tenants, eliminated attribute conflicts in the presence of assignments of attribute. This approach was believed to be accepted by various other types of trust beyond the presented trust types.

III. EXISTING SYSTEM

This section explains how the existing system is providing a secured access and its respective gaps. Security measures like Access control wherein the actions are limited on the resources acted by the user. Typically these policies are mentioned in the access control policies. Even though the given or suggested policies cover in multiple scenarios and own multiple features of expressions, and the complexity has some limitations for the practical usage. Indication of policy language conforms to all the data types and literal features packaged with the language. This manner ignores the truth that distinct demands of authorization on different web applications. Last few years, world has shown a lot of interest in environments which supports complex and different type of access control policies which has better features and expressiveness. Nowadays, when we think of most widely-accepted style standard for web service interfaces, certainly representational state transfer (REST) will come to our mind.

FLAWS OF EXISTING SYSTEM

- i) To accommodate certain systems, consumers must entirely need to have different types of platform specific policies.
- ii) Platform-specific authorization policy language had to be designed by Cloud Service providers on their own. They also need to take care of an efficient authentication and the related enforcement mechanisms.
- iii) Gaps in data storage security and auditing security in computations. The provider of cloud service have entire control of the customer's data, they can act on any malicious activities such as modify, destroy and copy the data. There are certain level of control mechanisms are followed on the virtual machines in cloud computing, because of this gaps in controlling the data which will eventually leads in lot of security issues compare to the generic cloud computing model.
- iv) Cheating of privacy
- v) Overhead performance computation and communication in-efficiencies.

IV. PROPOSED SYSTEM

To overcome the current gaps mentioned in the previous section, we have focused our research towards an already implemented access control named REST Policy Language (Rest-PL) which is typically request oriented. In practical usage and for better compatibility, scope reduction and its application to REST interfaces by RESTPL. The API and its respective multi-level authentication is designed based

on RestPL. The Jelastic server works based on the request oriented. Before the given request is processed and the result is displayed, multi-level authentication in terms of root signature verification, symmetric key authentication and file level access being performed. The request based feature ensures that based on actual request, automatic generation of RestPL policy which helps to avoid a user's designing a policy along with secured way of accessing the cloud services.

Phases:

- A. JELASTIC SERVER PHASE
- B. ROOT SIGNATURE GENERATION
- C. ROOT SIGNATURE THROUGH SMTP
- D. SECURE KEY THROUGH AES
- E. FILE LEVEL AUTHENTICATION
- F. FILE ACCESS-TIMELINE PRIVILEGES

A. JELASTIC SERVER PHASE

In the Netbeans development platform a plug-in created by Netbeans team reduces the application management process in the Jelastic platform. Prior to accessing the Jelastic cloud, mail id and password to be entered by the user to login into the Jelastic cloud. This id and password must have shared or sent to the user during the process of registration by the cloud provider. Post which installation of netbeans plug-in into the netbeans by the user. User will be able to gain the data access post matching their given credentials else the access will be denied.

B. ROOT SIGNATURE GENERATION

This module explains the contribution of owner part in Jelastic server. The authentication in terms of checking the user's mailid, password and finger print by the owner of the firm. Recognition of Fingerprint technique is a well-known technology and it is widely accepted in the entire world. If the given finger print has any kind of wet or dryness and dirty, it can result in inaccuracy. Sensors are used to capture the fingerprints and it provides a scanned image of the finger. A unique password generated as soon as the fingerprint is provided. Cloud service provider stores the Image and password in their CSP database.

Post registration when a user wants to access the cloud service, their scanned fingerprint image is sent to CSP database where the matching process is carried out with the already stored image. If the password details are matched against the registered one, then the process of splitting the image of a fingerprint into eight parts with the help of boundary splitting algorithm. Post splitting, input is given as eight parts to the merkle hash tree for the creation and generation

of root signature to the given fingerprint image. This is purely based on a technique called binary tree of hashes.

In the MHT, every leaf node is holding the hash of a data block, Internal nodes has the complete details / hash of the concatenated hashes of their children. This invented algorithm is one of the safe/trusted ways to share the root of the tree between the Verifier and Signer. For integrity verification of any data block, transmission of entire tree of hashes not necessary to the verifier. Authentication path of data block is alone to be considered when a signer transmits the hashes of only those nodes, then the sharing of root signature to the Jelastic server done by the owner and same is maintained or stored in the Jelastic server.

C. ROOT SIGNATURE THROUGH SMTP

Encrypted format of Root signature is stored in the Jelastic server using Advanced Encryption Standard. Further step involves the sending of root signature along with user name and password credentials to already registered mail id using Simple Mail Transfer Protocol (SMTP). User can access the server and their respective files when the DB root signature is matched with the given root signature.

D. SECURE KEY THROUGH AES

In this module, the managers of the project or HR or Finance or Engineering departments uploads their department specific files to the jelastic cloud, so that the service provider can allocate it to the respective team or a person. In this, Advanced Encryption Standard algorithm is used to encrypt the file. The files which are encrypted using the AES algorithm are saved into the database of Jelastic Cloud. Here, the managers can initiate for the secure key using the Symmetric key Encryption method and the same can be shared to the user.

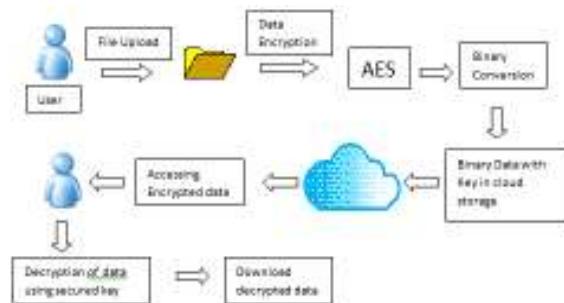


Fig 1. Pictorial representation of encryption and decryption mechanism

E. FILE LEVEL AUTHENTICATION

In the file level authentication process, user will receive the secure key shared by the administrator or the manager. User must enter the valid secure key prior to access the file stored in the Jelastic cloud. The secured key also to be entered for the process of accessing or decrypting the file.

F. FILE ACCESS-TIMELINE PRIVILEGES

Post successful dispensation to access the file, timelines are defined for each and every individual or to the team so that user can access the file within a defined timeline or specified duration. Revoke access is enabled when a user doesn't access the file for some time. Once the file access is revoked or access expires then the user will not be able to access to that particular file. In such cases, user to initiate a fresh request to the administrator for the extension of validating the timelines of the file. File access can be invoked/extend once the admin/owner accepts the request.

V. ARCHITECTURE DIAGRAM

Below is the basic framework and a high level structure of an application system. The structured solution has been defined to meet all the technical and operational requirements. This optimizes quality attributes such as security and manageability.

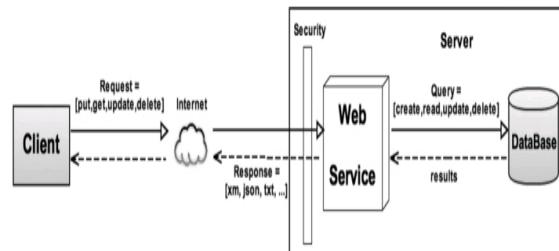


Fig 2. Secured communication flow between Client, Web service and Database.

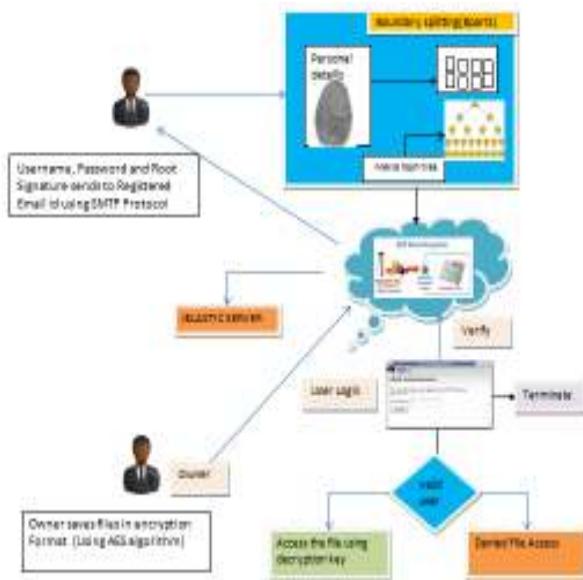


Fig.3 Architecture diagram on the authentication process of Jelastic Server

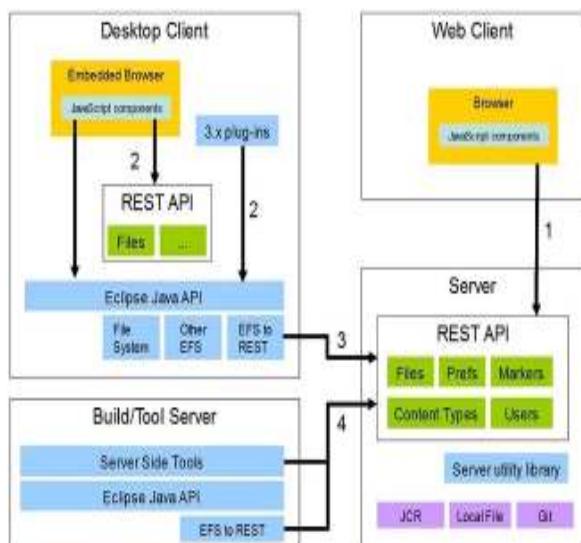


Fig.4. Layers of desktop, Server and Web-Client

VI. ADVANTAGES OF PROPOSED SYSTEM

- a. Uncheatable cloud computation and privacy cheating discouragement.

To define the security model in the cloud computing and to discourage the adversary from leaking the cloud user's sensitive data, we have introduced finger prints, root signature and secured key through AES mechanism as this will indicate the trust level of computation security and storage security, respectively. Above mechanism will ensure cloud computation or cloud storage as fully trusted.

- b. Reduction of computation cost
- c. Efficient data storage security by using multi- level authentication and validation
- d. Computation audit security

By the establishment of above said modules, Third Party Auditor (TPA) may concurrently handle multiple auditing delegations upon different users' requests. In addition to allowing TPA to perform the multiple auditing tasks concurrently, it also reducing the cost of computation from TPA side.

- e. Achieving communication efficiency using advanced protocol. For an effective communication, we have implemented a plaintext-equivalent authentication protocol and verifier-based authentication protocol

- f. Enhanced security.

VII. SECURITY ANALYSIS OF PROPOSED FRAMEWORK

FEATURES	PROPOSED SYSTEM	EXISTING SYSTEM
Secured Access management	Secured storage of all the credentials of the user done in Jelastic server. Jelastic Server validates the availability of unique ID for each user at the time of registration phase itself. Scanned fingerprint image is sent to CSP database where the matching process is carried out, then the process of splitting the image of a fingerprint into eight parts with the help of boundary splitting algorithm using Merkle Hash Tree concept.	Current system follows only the generic access control mechanism which has the basic security authentication system
Secured Credential Management	This ensures finger print image capture, root signature generation and transfer of secret key maintenance in a secured manner. For authentication purpose, registered or valid user's finger print is required. Timeline factor for the file level access is also an important feature in this framework. All these features makes the framework inherently stronger compared to the legacy credential management system	System deals only on the complex access control policies. RestPL is compatible with any access control tasks derived from a system that is running REST interfaces
Phishing attack	Based on multi-factor authentication, our model provides mutual authentication between the user and the Jelastic server. Apart from finger print authentication, Secret key, root signature is required for authentication. Moreover, timelines are defined for each and every individual or to the team so that user can access the file within a defined timeline. Hence, our framework efficiently defends the fishing attack	RestPL supports secured authorizations but the system not inclined to root signature, secret key and timeline privileges for file level access. Hence, it is likely vulnerable to phishing attack.
Man In The Middle Attack	In this, when an attacker tried unauthorized access into the system by using through the registered user ID and password, he/she will not be able to access the cloud services and resources because the user needs authentication which requires secret key generation which will be sent through their registered mail id. These credentials are exchanged between the server and a registered user using separate secured channel.	RESTPL system concentrates only on the policy as it can be automatically generated from an actual request, which helps mitigate a user's pressure from policy designing. The application did not specifically concentrate on the security related access.
REST Policy Features	The Proposed system is designed for REST with the concepts of approval along with multi factor secured authentication.	The existing system is designed for REST which has the concepts of approval and prohibition.

Table 1. Comparative analysis between proposed and existing system

VIII. RESULTS AND DISCUSSION

This paper portrayed unique finger impression recognition and multi-level secured access technique. In summary, to overcome the limitations of access control in RESTPL services as stated above in the same paper, I have designed an application which is based on request-oriented and multi factored secured access control which reduces the security issues faced by the cloud consumers and cloud service providers.

In addition to this RESTPL policy, additional security features are added to ensure data storage in terms of enhanced security, computation auditing security, achieving communication efficiency using advanced protocol.

Depicted below are the snapshots of outcome of the implementation of multi-level authentication mechanisms for cloud API in Jelastic server based on RestPL.



Fig 5. Registration and authentication check of a user



Fig 6. Login form of a user's screen

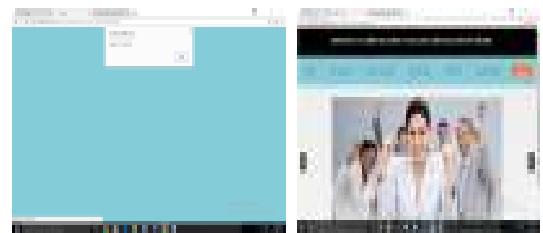


Fig 7. Login form of a user and home page

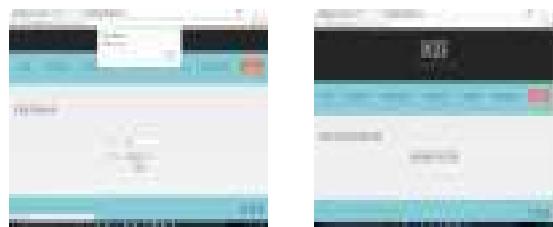


Fig 8. File upload and its respective details



Fig 9. File sharing and validity details



Fig 10. File receipt for the user

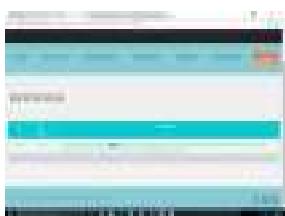


Fig 11. File access through secure key and receipt of files

IX. CONCLUSION

Non-matching rates of unique and identified mark matchers are growing on account of misshaped fingerprints. This produces a security gap in programmed unique mark acknowledgment frameworks which can be used by hoodlums and terrorists. Thus, it is important to add to a unique mark contortion identification and correction calculations to address the gap. This paper portrayed a unique finger impression recognition and multi-level secured access technique.

Future Work:

In summary, to overcome the limitations of access control in RESTPL services as stated above in the same paper, I have designed an application which is based on request-oriented and access control which reduces the complexities of cloud service provider and its consumers. The API and its respective authentication is designed based on RestPL. The Jelastic server works based on the request oriented.

Before the given request is processed and the result is displayed, multi-level authentication in terms of root signature verification, symmetric key authentication and file level access being performed. In addition to this policy, additional security features are added to attain data storage enhanced security, computation auditing security, achieving communication efficiency using advanced protocol.

The current research work involved the implementation of multi-level authentication using finger print image and maintain the security consistency till the end or exit of the access. Another impediment is that the present methodology does not bolster moved fingerprints. It is hard to gather numerous moved fingerprints with different twisting sorts and in the interim get precise mutilation fields for learning factual bending model. Further research work in the future will involve to address the above confinements.

Acknowledgment

This paper is specially intended to analyze and implement the best mechanism to overcome the current security and access related challenges through API, for this research I thank my guide Dr.P.Sujatha for her review and her motivational background helped to narrow down the analysis and research direction to the focused one.

References

- [1] Yang Luo, Hongbo Zhou, Qingni Shen, Anbang Ruan, Zhonghai Wu. "RestPL: Towards a Request-Oriented Policy Language for Arbitrary RESTful APIs". IEEE International Conference on Web Services, DOI 10.1109/ICWS.2016.92, June, 2016.
- [2] Stine Lomborg, Anja Bechmann. "Using APIs for Data Collection on Social Media". The Information Society, 30: 256–265, 2014, ISSN: 0197-2243 print / 1087-6537 online, DOI: 10.1080/01972243.2014.915276
- [3] Muhammad Imran Hussain and Naveed Dilber. "Restful web services security by using ASP.NET web API MVC based". Journal of Independent Studies and Research, Computing Vol. 12, Issue 1, 2014.
- [4] Jyoti Joshi. "Extended JV-RBAC Model with Secure API Access Control in Cloud". International Journal of Emerging Research in Management &Technology, Vol.4, Issue-6, June. 2015.
- [5] Navid Pustchi, Ravi Sandhu. "MT-ABAC: A Multi-Tenant Attribute-Based Access Control Model with Tenant Trust". Springer International Publishing Switzerland 2015, M. Qiu et al. (Eds.): NSS 2015, LNCS 9408, pp. 1–15, 2015, DOI: 10.1007/978-3-319-25645-0 14.
- [6] Aneta Poniszewska-Maranda. "Model Driven Architecture for Modeling of Logical Security Based on RBAC Approach". Journal of Applied Computer Science, Vol. 22 No. 1 (2014), pp. 183-199.
- [7] Kui Liu², Zhurong Zhou¹. "Towards a RBAC Workflow Model for Thesis Management". Journal of software, Vol. 10, No.4, April. 2015. DOI: 10.17706/jsw.10.4.480-490
- [8] Sonia Gupta, Rubal Choudary. "Multi Tenancy Access Control Using Cloud Service in MVC". IJEDR, ISSN: 2321-9939, Vol.3, Issue 4, 2015.
- [9] Adebukola Onashoga, Adebayo Abayomi-Alli, Timileyin Ogunseye. "Enhanced Role Based Access Control Mechanism for Electronic Examination System for Electronic Examination System". I.J.Computer Network and Information Security. DOI: 10.5815/ijcnis.2014.03.01, 2014
- [10] Zahid Pervaiz, Walid G. Aref, Nagabhushana Prabhu. "Accuracy-Constrained Privacy-Preserving ACM for relational data". IEEE Transactionson knowledge and data engineering, Vol. 26, 2014.
- [11] Pratik Bhingardeve, Prof. D. H. Kulkarni. "Security and Accuracy Constrained Task-Role based Access Control and Privacy Preserving Mechanism for Relational Data". International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 4 Issue 07, July. 2015.
- [12] Ting Cai, Jian Zheng, Xing Du. "A Hybrid Attribute based RBAC Model". International Journal of Security and Its Applications, Vol.9, No.7.2015, http://dx.doi.org/10.14257/ijisia.2015.9.7.29

Bitmap Indexes for Faster Query Execution

Samadhi P. Kumarasinghe, Nishadi D. Kirielle, Malmee Weerasinghe, Sasini Madhumali, Amal S. Perera,
Sachini Herath, Amila De Silva, Lasantha Fernando

*Department of Computer Science and Engineering
University of Moratuwa
Sri Lanka*

samadhipoornima.12@cse.mrt.ac.lk, ndimeshi.12@cse.mrt.ac.lk, wmalmee.12@cse.mrt.ac.lk, sasini.12@cse.mrt.ac.lk, shehan@cse.mrt.ac.lk,
sachini.h@cse.mrt.ac.lk, jaadds@gmail.com, lasantha.fdo@gmail.com

Abstract—

Existence of enormous amount of data and the necessity to extract useful information out of that data has made data analytics a profound topic today. Despite the many prevailing optimizations, one optimization that benefit certain types of queries might not benefit others. Hence understanding the proper technique is crucial. This paper focuses on effective usage of bitmap indexing to accelerate Online Analytical Processing (OLAP) queries. Queries that can be totally run on bitmap indexes is discussed in two different warehouses one which runs on conventional MapReduce paradigm and the other on In-Memory computational models. Queries that can narrow the search through bitmap index usage is discussed with respect to MapReduce paradigm. In addition, several enhancements that yield the advantage of fast bitwise logical operations is discussed.

Keywords—*Bitmap Indexes, Hive, Spark SQL*

I. INTRODUCTION

The analytical results gained by big data analysis has done a transformational advancement towards effective marketing strategies and much more benefits in businesses. Because the queries are often complex and the warehouse database is often very large, processing with minimal latency for interactive queries is a critical issue in the current data warehousing environment. The big data analytics platform discussed in our work focuses on improving efficiency of data analytics of large datasets thus addressing the above mentioned issue. The analysis is usually performed with queries that aggregate, filter, and group data in a variety of ways. One of the possible and best solutions for faster data processing is through the mechanism of indexing. Currently various indexing techniques exist, such as B-trees, Bitmap indexing, R*-tree indexing and p-trees. Each technique is suitable for a particular situation and thus might not be that effective when used for different situations.

This research paper discusses the effort to develop and evaluate a bit-oriented analytics platform (storage engine) designed to improve OLAP and other query-intensive applications. In particular, the paper discusses the type of queries and type of environments under which bitmap indexes are capable of improving query performance.

In our approach, we test for the performance of queries using bitmap indexes in two main contexts, the conventional MapReduce paradigm and the In-Memory computational models. Apache Hive is a data warehouse environment that uses MapReduce paradigm. Bitmap indexing is introduced on top of Apache Hive and used for queries that can totally run on indexes as well as queries that require base table access after the search is narrowed down using bitmap indexes. In addition, bitmap indexing is introduced on top of Apache Spark SQL and the performance of the queries that can be totally run on bitmap indexes is assessed.

The rest of the paper is organized as follows. In section II, the background of bitmap indexing and the motivation for the research is discussed. Section III contains the implementation details and the optimizations performed. It is followed by the performance evaluation carried out on the two Proof of Concepts (POC). In the end the results and possible future work are discussed along with the conclusion of the research.

II. BACKGROUND

In 2005, C-Store project brought out the concept of read optimized relational database management system [1]. This was a turning point in big data analytics with a considerable efficiency gained by column-wise storage of data. Commercializing the design of the C-Store research prototype, the Vertica analytic database was implemented providing a distributed and parallel RDBMS system [2]. Our research work takes a similar approach to what has been done in HP Vertica with column data, but instead of columns we use bitmaps.

Among the many researches done in the past to evaluate the effectiveness of indexing, in Ref. [3], bitmap indexing has been identified to be well suited for ad hoc queries when run on environments that involve large amounts of data. Ref. [4] points out that read-only data is more flexible in terms of indexing as opposed to dynamic data. Hence the existing literature suggested that the optimal bitmap index usage can be gained when used in a data warehousing environment with OLAP queries, which was our target domain as well.

Ref. [5] has concluded that the bitmap indexing is a time efficient approach when it comes to query execution due to fast binary operations performed at index level. Moreover, an investigation to select the proper indexing technique has been

carried out in Ref. [6] where 3 indexing techniques have been compared. This research included bitmap indexing as one of the three indexing techniques. The research showed that the major advantage of using bitmap indexing on a data warehouse is the ability to answer queries at the index level rather than going to the data level. In our research we extend this concept further. Our first step is investigating the type of queries that can totally run on bitmap indexes. Next, we analyze the performance in two different data warehousing environments to discover the optimal context for bitmap index usage.

Several bitmap indexing techniques have been discussed in Ref. [8] which are targeted for read most environments. Inspired by Ref. [8], we extend our research to assess the usage of bitmap indexes in different situations. Considering the existing bitmap compression techniques, Enhanced Word-Aligned Hybrid (EWAH) compression technique provides better performance in bitwise operations [10] and therefore is used in our implementation.

III. PROPOSED SOLUTION

In order to proceed with the concept of bit-oriented analytic platform, an existing database engine was selected as the base component. 304 Database engines were chosen initially and narrowed down considering a specific criteria for the bit-oriented analytic platform. Main considerations were ability to manage large sets of data and ability to run OLAP queries, existence or potential to implement bitmap indexes on top of the database, the licence and popularity. The mentioned criteria narrowed down the search to Hive and Spark, the former which makes use of the MapReduce paradigm and latter which is an In-Memory database engine.

Bitmap indexing was implemented in the project with support for the following;

- Bitmaps for values
Separate bitmaps can be created for each value in the given column.
- Bitmaps for ranges
Separate bitmaps for a given range out of the values of the mentioned column.
- Bitmaps for join indexes
When the queries run on the join of several tables, then the bitmap indexes are created for the column values of the joined table.

A. Hive

Apache Hive is a data warehouse infrastructure which incorporates the MapReduce paradigm. It is built on top of Hadoop and is targeted at data analysis on databases and file systems that integrate with Hadoop.

1) Existing bitmap Indexes on Hive.

Indexing was added in Hive in version 0.7.0, and bitmap indexing was added in version 0.8.0. Hive uses the javaEWAH library which is a project by Daniel Lemire, for

the bitmap index creation and bitwise operations. This library takes the position of 1's in the bitmap as input and provides the compressed bitmap as output.

The bitmap index table consists of four columns namely the columns that are being indexed, bucketname, offset, and bitmaps. Here, the bitmaps column contains the bitmaps created on the row offset. However, for data stored in certain formats including text format, there is only one entry per row. Therefore, the row offset of a particular entry always remains 0. Hence the bitmap index for all the entries remain the same. This has both storage and performance drawbacks. This implementation results in the index table having same number of entries as the base table thus hindering the possible storage advantage of using bitmaps. It also restricts the effective usage of bitmap indexes which is explained in detail in the next paragraph.

Index usage of Hive happens in two stages. In the first stage it looks for all the entries that satisfies the criteria required in the query. This is done by joining the relevant index tables. This is a normal table join with the additional condition that the result from AND or OR operations of the relevant bitmaps should not be empty. Since the bitmaps are always the same, these logical operations on them give out the same bitmap thus resulting it never being empty. Hence the role bitmap indexes play in the query execution is of no advantage. Next, the bucketname and the bucket offset pairs of entries that satisfy the query criteria are written into a temporary directory to be accessed in the second iteration. In the second stage of the query it visits only these filtered locations and extract required data from those locations. In this way it avoids the necessity to run through all the entries in the table thus reducing the query execution time.

2) Bitmap Index Creation.

According to the existing bitmap indexing in Hive, the role the bitmap indexes play when executing a query was not adequate to gain the advantage of fast bitwise logic operations on them. Hence, a new implementation of bitmap indexing was introduced to investigate the research topic at hand. This implementation differed to original bitmap index implementation on Hive in that instead of row offsets, here we created bitmaps for block offsets.

Now, the bitmap table only contains three columns namely the columns that are being indexed, bucketname, and bitmap for block offset. By taking the block offset instead of row offset, we gained the following two advantages.

1) Each row contains a single table entry. Since each block contain a set of such rows, the block offset of a row is unique within the block. And then by grouping all the offsets that contains the same value in that block, we can create an index unique to each distinct value. Therefore, in the new bitmap index table we have a row per distinct values only. This gains a storage advantage over the original bitmap index implementation where there was a row per each value.

2) Since the block offset is unique to each row in a column, we can perform AND and OR operations on the bitmaps and decode the offsets from the resultant bitmap in query execution. However, since the current implementation creates bitmaps within blocks, when the table spreads over several blocks, it is important to make sure that the bitwise operations are carried out blockwise.

As further improvements, we introduced following two features for bitmap index usage.

1) Using “UNION ALL” instead of “JOIN”

Usually an AND/OR operation takes two arguments. Hence in situations where we want to run such a logical operation on more than two bitmaps, we have to perform it on two bitmaps and then take the resultant bitmap and perform the logical operation on it and so on. When it comes to bitmap index tables, we have to join two bitmap index tables, take the result and perform a join with the next bitmap index table and so on. But if we can have a mechanism to take all the bitmaps and perform AND/OR operation on them all at once we can replace multiple usage of expensive JOIN operations with a single instance of UNION ALL. To incorporate this, we introduced two user defined functions(UDF) COMPOUND_AND and COMPOUND_OR, which take a list of bitmaps and AND/OR all the bitmaps in the list and return the resultant bitmap.

COMPOUND_OR also has some other advantage in queries with “between” clauses. For range queries to run on bitmap indexes, we should know in advance all the values within the range and perform multiple OR operations on them or we should create a bitmap index for the range. But now we can simply collect all the bitmaps within the given range into a list and send it to the COMPOUND_OR without having to know the exact values within the range.

2) Carrying out COUNT queries solely on bitmap indexes.

JavaEWAH provides the functionality to obtain the count of 1's in the bitmap. This functionality was incorporated and a new UDF (BITMAP_CARDINALITY) was introduced to carry out count queries directly from the indexes.

B. Spark SQL

Spark SQL provides better advantages with In-Memory computational model compared with other existing paradigms. Although indexing techniques enhance query performance in traditional database systems, index structure becomes a performance bottleneck for In-Memory databases. Hash tables come in handy in these situations but as they support only for the point queries, they can not be used. In order to overcome these problems, at the initial stage, we focused mainly on the goal of improving the performance of the queries that can be run solely using bitmap indexes. Indexing is not included Spark SQL, thus bitmap indexing was introduced as a part of our implementation.

The major advantage of bitmap index usage here is that, the system need not load all the data to memory for processing. Instead, it can load only the relevant bitmap indexes. As the bitmaps are compact and compressed, this gives a storage efficiency and reduced execution time due to lesser data loading time.

1) Bitmap Index Creation.

The user is given the ability to create indexes on the datasets with a separate Spark programme. The bitmaps thus generated are inserted to a table and saved in a separate file.

2) Incorporating Bitmap Indexes in Query Execution.

The bitmaps are stored in the EWAH compressed bitmap format and the processing is done using the bitwise operations provided by the JavaEWAH library. In order to incorporate this functionality to the Apache Spark SQL platform, we have introduced new UDFs and registered them in the functions registry. The functions that can be used to incorporate bitwise operations are defined in Table I. These UDFs are then optimized using the Spark inbuilt catalyst optimizer.

TABLE I. USER DEFINED FUNCTIONS ON SPARK SQL

Function Name	Functionality	Input parameters	Return value
BITMAP_CARDINALITY	Calculates the number of 1's in the bitmap	An EWAH-compressed bitmap	The counted value of 1's in the bitmap
BITMAP_AND	Performs the bitwise AND operation between the two bitmaps	Two EWAH compressed bitmaps	The resulting EWAH compressed bitmap
BITMAP_OR	Performs the bitwise OR operation between the two bitmaps	Two EWAH compressed bitmaps	The resulting EWAH compressed bitmap

IV. PERFORMANCE EVALUATION

TPC-DS dataset has been used for the purpose of performance evaluation of the two POCs. TPC-DS is the main industry standard benchmark for measuring the performance of decision support solutions for Big Data systems.

A. POC on Apache Hive

The following results has been tested on hortonworks-Hive testbench for TPC-DS 1GB dataset with a Apache Hive standalone cluster. In this dataset, each table

fitted into one block. However, when moving on the larger data which expands over multiple blocks, rewritten query should include a “WHERE” clause to consider the blockId prior to carrying out bitwise logical operations or increase the block size so that the data fit in a single block. The indexes have been created on text data whereas its performance has been tested against Optimized Row Columnar(ORC) formatted dataset, which has a higher performance compared to the rest of the file formats.

Queries that can totally run on bitmap-indexes.

The query 96 of TPC-DS query set, which is defined in Table II is an instance where we can run the whole query on top of the bitmaps. Fig. 1. Shows the execution times obtained for query 96. Here, it includes the execution times when the query was run on Hive without any index usage (first column), Hive with the existing bitmap index usage (second column) and the newly introduced bitmap indexing technique which is in the column named “bit-store” (third column).

TABLE II. QUERY 96 - ORIGINAL AND RE-WRITTEN QUERIES

Original Query	Rewritten Query
<pre>SELECT count(*) as c FROM store_sales, household_demographic s ,time_dim, store WHERE store_sales.ss_sold_time_sk = time_dim.t_time_sk and store_sales.ss_hdemo_sk = household_demographic s.hd_demo_sk and store_sales.ss_store_sk = store.s_store_sk and time_dim.t_hour = 8 and time_dim.t_minute >= 30 and household_demographic s.hd_dep_count = 5 and store.s_store_name = 'ese' order by c limit 100;</pre>	<pre>SELECT BITMAP_CARDINALITY(COMP OUND_AND(COLLECT_LIST(d. bitmaps))) as e FROM (SELECT `_bitmaps` AS bitmaps FROM tpcds_orc_1_table96_t_hour_proj WHERE t_hour = 8 UNION ALL SELECT `_bitmaps` AS bitmaps FROM tpcds_orc_1_table96_t_minute_p roj WHERE t_minute_greater_equal_30=TRUE UNION ALL SELECT `_bitmaps` AS bitmaps FROM tpcds_orc_1_table96_hd_dep_co unt_proj WHERE hd_dep_count=5 UNION ALL SELECT `_bitmaps` AS bitmaps FROM tpcds_orc_1_table96_s_store_na me_proj WHERE s_store_name = 'ese')d order by e limit 100;</pre>

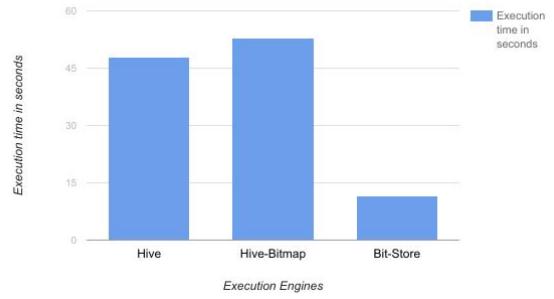


Fig. 1. Execution times for TPC-DS query 96

In this query, it can be seen that Hive without indexes has a lesser execution time compared with bitmap indexed Hive. That is because in the conventional way Hive uses indexes on their queries, first it narrows down the search using indexes and then visit the actual locations. But for a “COUNT” query, revisiting the actual locations is not necessary. That redundant work increases the execution time. This is avoided in bit-store since the count is done directly on the bitmap indexes thus avoiding the temporary read and write. Hence the newly introduced bitmap indexing has higher performance.

Queries that require accessing the base table.

Execution of these type of queries happen in two stages. The first stage is to obtain the locations of the data relevant to the query so that when the actual execution happens, visiting only the selected locations is sufficient. The obtained locations are written to a temporary directory for access in the second stage. Next we run the query to execute on the locations filtered. Note: According to the way Hive has implemented index usage, it only facilitates the index usage for queries that involve only one MapReduce job. Hence, the “order by” clause in the original query cannot be executed with index usage in the current context. Therefore, the performance evaluation has been carried out excluding the “order by” clause in both original and rewritten queries. Table III indicates query 42.

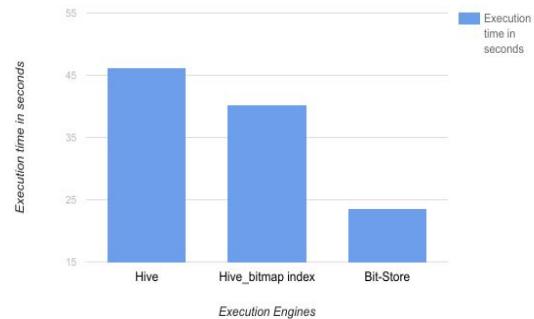


Fig. 2. Execution times for TPC-DS query 42

TABLE III. QUERY 42 - ORIGINAL AND RE-WRITTEN QUERIES

Original Query	Rewritten Query
<pre> SELECT dt.d_year, item.i_category_id, item.i_category, sum(ss_ext_sales_price) as s FROM date_dim dt ,store_sales,item WHERE dt.d_date_sk = store_sales.ss_sold_date_sk and store_sales.ss_item_sk = item.i_item_sk and item.i_manager_id = 1 and dt.d_moy=12 and dt.d_year=1998 group by dt.d_year ,item.i_category_id ,item.i_category order by s desc,dt.d_year, item.i_category_id, item.i_category limit 100 ; </pre>	<pre> INSERT OVERWRITE DIRECTORY "/tmp/index_result" SELECT b.bucketname AS `_bucketname`, e.offset as `_offsets` FROM (SELECT BITMAP_POSITIONS(COMPOUND_AND(COLLECT_LIST(d. bitmaps))) as offset FROM (SELECT `_bitmaps` AS bitmaps FROM tpcds_orc_1_table42_i_manager_id_proj WHERE i_manager_id = 1 UNION ALL SELECT `_bitmaps` AS bitmaps FROM tpcds_orc_1_table42_d_moy_proj WHERE d_moy=12 UNION ALL SELECT `_bitmaps` AS bitmaps FROM tpcds_orc_1_table42_d_year_proj WHERE d_year=1998) dje, (SELECT DISTINCT `_bucketname` AS bucketname FROM tpcds_orc_1_table42_d_year_proj)b GROUP BY bucketname, offset; </pre>

Fig. 3 indicates the query execution times obtained for three queries TPCDS query 3, TPCDS query 52 and TPCDS query 55. These queries were run in a similar manner to the TPCDS query 42. Query 3 requires two AND operations here as the query 52 and 55 requires three AND operations. Since the bitwise operations are fast, queries which have higher number of AND/ORs gain a higher advantage.

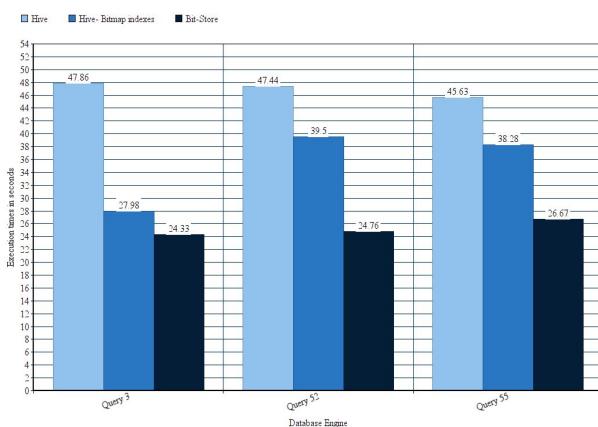


Fig. 3. Execution times for TPC-DS query 3, 52 and 55

B. Spark SQL

The query 96 in the TPC-DS query set is an instance where we can run the whole query on top of the bitmaps. The following results have been tested with Apache Spark standalone cluster with 4G executor memory. Query 96 is rewritten as in Table IV to be run on bitmaps.

In this rewritten query, the bitmap join indexes are used which are created on the columns of the join table of store_sales, household_demographics, time_dim and store. The bitmap indexing is used for the columns t_hour, s_store_name, hd_dep_count. For the t_minute column, as the user is interested in accessing the bitmap range indexes, the aggregate bitmap range indexes are created in advance and accessed in the query execution. The resulting execution times are indicated in Fig. 4.

TABLE IV. QUERY 96 - ORIGINAL AND RE-WRITTEN QUERIES

Original Query	Rewritten Query
<pre> SELECT dt.d_year,item.i_categ ory_id ,item.i_category ,sum(ss_ext_sales_pri ce) as s FROM date_dim dt ,store_sales,item WHERE dt.d_date_sk = store_sales.ss_sold_d ate_sk and store_sales.ss_item_s k = item.i_item_sk and item.i_manager_id = 1 and dt.d_moy=12 and dt.d_year=1998 group by dt.d_year, item.i_category_id ,item.i_category order by s desc,dt.d_year, item.i_category_id, item.i_category limit 100 ; </pre>	<pre> WITH bitmap1 AS (select bitmap from t_hour_index where value = 8), bitmap2 AS (select bitmap from s_store_name_index where value ='ese'), bitmap3 AS (select bitmap from hd_dep_count_index where value = 5), bitmap4 AS (select bitmap from t_minute_index where value = '>=30') SELECT BITMAP_CARDINALITY(BITMAP_AND(BITMAP_AND_T ERNARY(bitmap1.bitmap, bitmap3.bitmap, bitmap2.bitmap), bitmap4.bitmap)) FROM bitmap1,bitmap2,bitmap3, bitmap4 </pre>

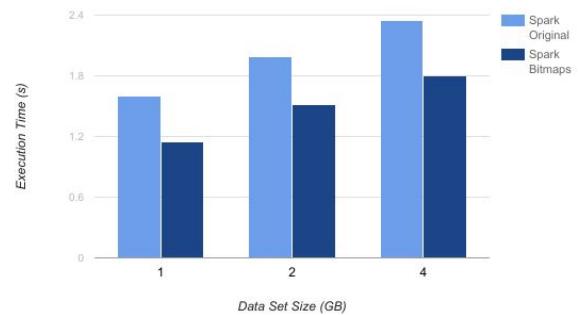


Fig. 4. Spark SQL TPCDS Query 96

V. DISCUSSION

In a conventional MapReduce paradigm, the number of map jobs and reduce jobs can be reduced by running the queries on indexes in situations where such execution is possible. In other situations, bitmap indexes can be used to narrow down the search so that running through the whole table is no longer necessary. Further, the performance gain increases as the “WHERE” clause of the query expands. That is because with a larger set of “WHERE” conditions, the number of bitwise logical operations in the rewritten query also becomes larger thus earning a greater advantage from the fast logical operations.

Several UDFs were introduced in the project as optimization aids. This shows the prospect of more UDFs (e.g. a UDF for XOR logical operation) targeting further optimizations. Moreover, the query rewrites can be automated using a set of rule-based syntax as a user experience enhancement.

In-Memory computational models like Spark SQL do not support indexing. These databases make use of the advantage of faster In-Memory processing rather than through indexing. However, by using a space efficient indexing mechanism like bitmap indexing, the amount of data that has to be loaded to memory for processing reduces. This is appropriate in situations where the query can be run solely on bitmap indexes. To gain the advantage of bitmap indexing by narrowing down the search and visiting only the filtered locations, bitmap indexing should be introduced to match the locations where data is stored. This remains as a possible future enhancement.

Bitmap indexing can be beneficial in different environments with the proper implementation and usage. In this paper, we have discussed several beneficial cases. However, to explore additional enhancements to bitmap index usage remains an interesting research area.

VI. CONCLUSION

Bitmap index usage proved to be beneficial in both MapReduce paradigms and In-Memory computational models. The queries that can be executed only on bitmap indexes earn the advantage of not having to access large base tables. Other queries that support index usage earn the advantage of narrowing the base table locations to be accessed. Moreover, bitmap index usage possesses the advantage of fast bitwise logical operations. Hence, proper usage of bitmap indexes is a productive read optimization in data warehouses for OLAP queries.

REFERENCES

- [1] M. Stonebraker et al., “C-store: a column-oriented DBMS,” in Proceedings of the 31st international conference on Very large data bases, 2005, pp. 553–564.
- [2] A. Lamb et al., “The Vertica analytic database: C-store 7 years later,” Proceedings of the *VLDB Endowment*, vol. 5, no. 12, pp. 1790–1801, 2012.
- [3] K. Stockinger, K. Wu and A. Shoshani, “Strategies for processing ad hoc queries on large data warehouses..,” in Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP, ACM, pp. 72–79.
- [4] G. Powell, Beginning database design. John Wiley & Sons, 2006.
- [5] W. Weahama, S. Vanichayobon, and J. Manfuekphan, “Using Data Clustering to Optimize Scatter Bitmap Index for Membership Queries,” in International Conference on Computer and Automation Engineering, 2009. ICACE ’09, 2009, pp. 174–178.
- [6] S. Jamil and R. Ibrahim, “Performance analysis of indexing techniques in Data warehousing,” in Emerging Technologies, 2009. ICET 2009. International Conference on, 2009, pp. 57–61.
- [7] P. O’Neil and D. Quass, “Improved query performance with variant indexes,” in *ACM Sigmod Record*, 1997, vol. 26, pp. 38–49.
- [8] D. Lemire, O. Kaser and K. Aouiche, “Sorting improves word-aligned bitmap indexes” *Data & Knowledge Engineering*, pp.3-2

Optical Flow For Robot Navigation

Nixon Adu-Boahen

Department of Computer Science
KNUST
Kumasi, Ghana
nix9281@gmail.com

J. B. Hayfron-Acquah

Department of Computer Science
KNUST
Kumasi, Ghana
jbha@yahoo.com

J. K. Panford

Department of Computer Science
KNUST
Kumasi, Ghana
jpanford@yahoo.com

Abstract — In this paper optical flow algorithm was implemented in robot to establish optical flow's efficiency in robot navigations. It was found that, the computation methods designed earlier (at the inception period) have also been improved tremendously by several researchers over decades now, but they are rarely implemented practically. In this work, robot navigations were experimented using web-cam with frame dimension of 320 X 240 as robot's eye with fundamental optical flow (precisely Lukas-Kanade which is one of the fundamental optical flow computation methods) results as navigation parameters. The subject was able to perform navigation using the optical flow results as parameters for navigation. Demonstrations perform disclosed that robots can easily navigate based on less expensive vision sensors such as webcam or CCD cameras with optical flow algorithms instead of sonars or radars.

Keywords-component; *Optical flow; robot; navigation; efficiency; vision; webcam, metaheuristics*

I. INTRODUCTION

Computer Vision is one of the fastest growing fields in Artificial Intelligence (AI). One of the greatest challenge met by this field is Robot Navigation. To submerge the challenge in robot navigation, researchers came up with many methods and techniques. Those methods can be grouped into vision and non-vision based. The non-vision based applies sound and light technologies such as Radar, Lidar and Sonar to detect objects in the robot's environment. The vision based methods include laser-light and optical flow. Optical flow is one of the recent vital subjects of interest triggered by autonomous navigation of robots whose measurement is an early vision processing step in computer vision, which is used in a wide variety of applications, ranging from three dimensional scene analyses to video compression and experimental physics. After reviewing recent papers on robot navigation and optical flow, it was realize that, optical flow is still rarely used in robot navigations. This research has established that, optical flow has greatest potential in application of robot navigation and hence system designers in the AI can rely on the use of optical flow for robot navigations since it is efficient.

II. ROBOT NAVIGATION STRATEGIES AND ALGORITHMS

Robot navigation is an embodiment of all the things a robot needs to be able to move its entire or part of its makeup from

point A to point B as efficient as possible without bumping into an obstacle (such as furniture, walls or people). Such activities include knowing where it is (Localization), detecting and avoiding obstacles (Collision Avoidance), memorizing its surroundings (Mapping), being able to plan a route (Trajectory Planning) to point B, and being able to explore (Exploration) new terrain. To achieve the goals of navigation, strategies have been devised and researched into. Example of such strategies are balancing and time to contact estimation which gives urge to the robot for it to avoid obstacles or change properties of motion to get to its destination.

Robot navigation can be put into two major categories: known and unknown environment. With the known, the robot may have built-in map models with which it compares the searched environment during navigations whereas in the unknown, the robot may not have prior knowledge of the environment it navigates in. According to g [1], algorithms that are used to perform navigations in the unknown environment is termed as Bug Algorithms. Ng [1], after experimenting the performances of about eleven different Bug algorithms developed SensorBug algorithm, which reduces the frequency at which data about visible environment is gathered and the amount of scanning for each time data is gathered. Worrall [2] embarked on application of algorithms in cooperate robots to perform urban search and rescue. He describes search algorithm as the algorithm for finding spots in surroundings relevant to geolocation.

Prior to Worrall (2008) existed six robot searching algorithms. These were exhaustive, random, hillclimbing, Tabu, SA and GA. The Exhaustive algorithm also known as brute force is the basic or simplest form of the search algorithms which is used where techniques to solving a challenge is limited. Worrall [2] mentioned that, the exhaustive search has seen little literature due to its empirical nature. Random algorithm simply chooses solutions at random and tests the solutions until a stop condition is met. It was mentioned that, research regarding path planning based on the random algorithm has seen little literature. The Worrall [2] claimed that, more attention is given to HillClimbing's algorithm than the former two. The Hillclimbing algorithm had serious problems of returning a local optima. Warrall [2] said, 'Russell & Norvig (1995) moved for the motion that changes be made in the usual HillClimbing algorithm to erratically opt another tool where optimal solution of a sort is detected instead of returning local optima. The Tabu is portrayed as metaheuristic as it is intended to keep running in

help of other calculation once in the past depicted and direct its pursuit [2]. Despite the fact that the utilization of the Tabu calculation to accomplice an essential calculation takes up memory to permit the Tabu rundown (which is a rundown used to keep up a rundown of arrangements that have just been assessed), its favorable position is the capacity to defeat the nearby optima union issue [3]. The Simulated Annealing (SA) calculation is one of a kind which impersonates toughening (i.e. the procedure in which a fluid is cooled until the point when it ends up plainly stable in a strong frame). Worrall [2] said, the SA algorithm is inexactly identified with that of HillClimbing as it does a neighborhood seek. As indicated by McGookin & Murray-Smith [4], the SA has been utilized to advance controllers for marine vessels. Aside the controllers' application, it has likewise been connected to take care of numerous improvement issues, for example, the voyaging businessperson issue and has additionally been connected in the minimization of energy utilization in remote correspondence, Montemanni [5]. Genetic Algorithms (GA) are associated with Theory of Evolution by Charles Darwin, Ellis [6] in that it mimics the natural evolution. To perform GA, a normal series operators (selection, crossover and mutation) is selected. The examples of operator procedures include Roulette wheel, Tournaments, Ranking and Elitist (Selector methods), uniform, one point, two point and multi point (Crossover methods) and Genetic Farming (Mutation methods). The procedures as stated, are popular ones nonetheless every years there are updates as to new findings by researchers, Worrall [2]

As seen in the introduction, computer vision plays very important role in artificial intelligence as a whole. Many methods have been devised to help in achieving the goals of computer vision since its inception. For instance, devices such as ultra sound sensors and radars were developed in the early days for machine and computer vision, though still in use. Tompkin [7] said, "The published work of Horn and Schunck on optical flow triggered the rigorous searches into all kinds of motion estimation based optical flow". According to McCarthy [8], optical flow is the measure of visual motion induced by the movement of surfaces in a scene with respect to the camera. Optical flow concept was first studied in the 1940s and ultimately published by American psychologist [9]as part of his theory of affordance. Optical flow is one of the current key fields contributing to computer vision. Many papers have been published on optical flow since the inception of its studies; literally over thousand papers have been published on it since the last thirty years as mentioned [10]. Many of the works done on optical flow seeks to have the element of estimating motion based on optical flow which has very important role to play in computer vision [11]. Although several researches have been conducted on optical flow its practical implementation in robots for vision seem difficult; in fact the practical difficulty of robot's implementation pushed some researches to make the following comment, "in most cases, traditional general purpose processors and sequentially executed software cannot compute optical flow in real time" [12].

III. OPTICAL FLOW ESTIMATION METHODS

It is said that "Inventive work on computing image velocity for compressing TV signals dates back to the mid 70's [13]. During the 80's, the fundamental assumptions enabling optical flow estimation, namely, brightness conservation and flow field coherence, were examined from different angles resulting in a large number of techniques, which are compared in the influential review articles by Barron, Fleet and Beauchemin". Almost all the works in image sequence processing begins by attempting to find the vector field which describes how the image is changing with time. Ideally, the projection into the two dimensional image planes of the three dimensional velocity field seen by the camera should be computed which make it difficult in practice but it is a necessity. Currently there are four categories of the methods for estimating optical flow: phase correlation, block-based, differential and discrete optimization methods

A. Phase Correlation Method:

This is a way to deal with, appraise the relative translative counterbalance between two comparable pictures (advanced picture relationship) or other informational indexes [14]. It is explained that, the computational method for phase correlation is quite simple and is based on Fourier shift property and it states a shift in the coordinate frames shift of two functions is transformed in the Fourier domain as linear phase differences [15]. Phase correlation is commonly used in image registration and relies on a frequency-domain representation of the data, usually calculated by fast Fourier transforms. This term is applied particularly to a subset of cross-correlation techniques that isolates the phase information from the Fourier-space representation of the cross-correlogram.

• Benefits of Phase Correlation

It is relatively (compared to other methods) insensitive to illumination changes because it just makes use of phase information of cross-power spectrum [16]. Also this method can be extended to determine rotation and scaling differences between two images by first converting the images to log-polar coordinates. Due to properties of the Fourier transform, the rotation and scaling parameters can be determined in a manner invariant to translation [17].

• Limitations of Phase Correlation

This algorithm can only implement integer-pixel motion estimation hence highly accurate estimation cannot be achieved [16].

B. Block-Based Method

This method is an improvement on the phase correlation method. It adapts the reduction of squared difference summation, or maximization of normalized cross-correlation. For each overlapping window, the method computes the correlation score for each integer translational movement. These computations

can be executed in the frequency domain for efficiency reasons [18].

- Benefits of Block-Based methods

It provides improvement in accuracy of optical flow estimation. It is said that the block-based method brings deformation of the windows for computation and that produces significant improved results [18].

- Limitations of Block-Based Methods

The block-based method has relative high computational cost [18].

C. Differential Methods

The differential techniques for evaluating optical stream flow, bases on partial derivative of the image flag or signal and additionally the look for stream field and higher-arrange fractional subsidiaries. The following are examples of such differential algorithms: Horn–Schunck method which reduces proportion in relation to residuals from brightness constancy constraint as well as a particular normalization span expressed as expected smoothness of the flow field, Lucas–Kanade method which improves upon Horn-Shunck technique based on image patches providing an affine model for the flow field, Buxton–Buxton method which based on a model of the motion of edges in image sequences Humphreys & Bruce [19], Black–Jepson method which makes use of coarse optical flow via correlation [20] and the General variational methods which include a range of modifications/extensions of Horn–Schunck, using other data terms and other smoothness terms.

- Benefits of Differential methods

The local techniques such as (Lucas-Kanade) of the differential method offer relatively high robustness under noise. The global techniques such as (Horn–Schunck) also yield flow fields with 100% density [21].

- Limitations of Differential Methods

Until the year 2012, the local technique of the differential method had the limitation of generating low flow density though it was a very robust technique. Also, the global technique of the differential method had the drawback of being sensitive to noise although it yield a flow fields with 100% density [21]. The combination of the local techniques and global techniques by Bruhn, et al. had abridged the limitations which were associated with the individual techniques by juxtaposing the advantages of each technique (i.e. local and global). So there is no known limitations associated with this method of Optical flow computation except the general perceived problem of high computational cost of optical flow as a whole.

D. Discrete Optimization Methods

In discrete optimization method, the search space is quantized, and then image matching is addressed through label assignment at every pixel, such that the corresponding deformation minimizes the distance between the source and the target image

[22]. The optimal solution is often recovered through min-cut max-flow algorithms, linear programming or propagation methods.

- Benefits of Discrete optimization methods

The discrete optimization method reduces computational complexity and hence has high rate of accuracy than other optical flow methods [23].

- Limitation Discrete optimization methods

The discrete optimization methods is associated with an inherent sampling inefficiency [23] due to the required extra storage for the labelling which in turn makes it slower compared to other methods.

IV. GENERAL OPTICAL FLOW ALGORITHM AND APPLICATIONS

A. General Optical Flow Algorithm

The experiments for this work was based on one of the fundamental optical flow computational methods. The goal of optical flow estimation is to compute an approximation to the motion field from time-varying image intensity [24]. To perform the computations, 3-D images are projected onto 2-D planes and the corresponding 2-D vectors are extracted and analyzed to find the vector motion of the image within the plane. It has been said that, 2D / 3D derivatives generally are calculated via repetitive solicitation of lower and higher pass filters [25]. It was realized from their work that, computing optical flow based on differential (Lukas and Kanade algorithm) method involved two steps:

- Calculate and compute spatio-temporal intensity derivatives comparable to calculating normal velocities to the local intensity structures and
- Integrate normal velocities into full velocities, for instance, either locally via a least squares calculation or globally via a regularization.

To ensure that optical flow truly assumes real motions in scenes instead of expansions, contractions, deformations and shears of various scene objects, three assumptions were also made as follows:

- ✓ No occlusion that is one object moving affront or behind another object, unless modeled for.
- ✓ No secularities in scenes otherwise the light source(s) and sensor(s) positions needs modeling explicitly.
- ✓ Objects in the scene are assumed rigid and free from motion or altering.

Like other differential optical flow estimation Barron & Thacker estimated the motion constraint equation for 2D and 3D as follows:

Assume $I(x, y, t)$ is the center pixel in a $n \times n$ neighbourhood and moves by $\delta x, \delta y$ in time δt to $I(x + \delta x, y + \delta y, t + \delta t)$. Since $I(x, y, t)$ and $I(x + \delta x, y + \delta y, t + \delta t)$ are the images of the same point (and therefore the same) so we have:

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t). \quad (1)$$

The assumption forms the basis of the 2D Motion Constraint Equation and is illustrated in figure 1. The Assumption is true to a first order approximation (small local translations) provided $\delta x, \delta y, \delta t$ are not too big. A first Taylor series expansion about $I(x, y, t)$ is performed in equation (1) to obtain:

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) + \partial I \partial x \delta x + \partial I \partial y \delta y + \partial I \partial t \delta t + H.O.T. \quad (2)$$

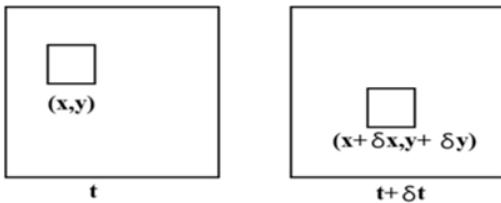


Figure 1 image at point (x, y, t) is same as at point $(x + \delta x, y + \delta y, t + \delta t)$

Where H.O.T. are Higher Order Terms, and assumed as small and could be ignored. Using equations (1) and (2) a new equation (3) is obtained:

$$\partial I \partial x \delta x + \partial I \partial y \delta y + \partial I \partial t \delta t + \partial I \partial x \delta x + \partial I \partial y \delta y + \partial I \partial t \delta t | \{z\} = 1 = 0 \text{ and finally :}$$

$$\partial I \partial x v_x + \partial I \partial y v_y + \partial I \partial t = 0. \quad (3)$$

Here $v_x = \delta x / \delta t$ and $v_y = \delta y / \delta t$ are the x and y components of image velocity or optical flow and $\partial I / \partial x$, $\partial I / \partial y$ and $\partial I / \partial t$ are image intensity derivatives at (x, y, t) . We normally write these partial derivatives as:

$$Ix = \partial I / \partial x, Iy = \partial I / \partial y \text{ and } It = \partial I / \partial t \quad (4)$$

Where $\nabla I = (Ix, Iy)$ is the spatial intensity gradient and $\vec{v} = (v_x, v_y)$ is the image velocity or optical flow at pixel (x, y) at time t . $\nabla I \cdot \vec{v} = -It$ is called the 2D Motion Constraint Equation and is 1 equation in 2 unknown (a line) as shown in Figure 2b. This is a consequence of the aperture problem: there is usually insufficient local image intensity structure to measure full image velocity but sufficient structure to measure the component normal to the local intensity structure.

Bear in mind v_x, v_y differentials are x and y optical flow components and (Ix, Iy, It) which are intensity derivatives. This equation can be rewritten more compactly as:

$$(Ix, Iy) \cdot (v_x, v_y) = -It \quad (5)$$

or as:

$$(\nabla I \cdot \vec{v}) = -It \quad (6)$$

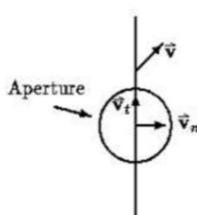


Figure 2(a) Aperture

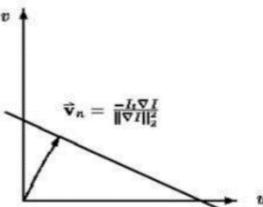


Figure 3(b) Normal Velocity

Figure 2(a) depicts an instance of problem of aperture, a moving line up and right is realized over circular aperture. In this case, it turns out to be very hopeless the right full picture speed, in any case the picture speed stays ordinary to the line. The full image speed computation issue at that point end up being finding an extra limitation that yields a moment different condition in similar questions. Ordinary speed at that point is a neighborhood marvel and happens when there is insufficient nearby power structure to enable a full picture speed to be recouped. For this situation, just the segment of speed ordinary to the neighborhood force structure (for instance, an edge), \vec{v}_n , can be recuperated. The extraneous part of the speed, \vec{v}_t , can't be recuperated.

In Figure 2(b), the 2D Motion Constraint Equation produces a line in $\vec{v} = (v_x, v_y)$ spectrum. One speed on this line is the right speed. The speed with the littlest extent on that line is the ordinary speed v_n . The greatness and bearing of the typical speed, $\vec{v}_n = v_n \hat{n}$ can be registered exclusively as far as the force subsidiaries, Ix, Iy and It as:

$$v_n = \frac{-It}{\|\nabla I\|_2} \text{ and } v_n = \frac{(Ix, Iy)}{\|\nabla I\|_2}. \quad (7)$$

v_n and \hat{n} are the raw normal velocity magnitude and the raw normal velocity unit direction respectively, i.e. :

$$\vec{v}_n = v_n \cdot \hat{n} = \frac{-It(Ix, Iy)}{\|\nabla I\|_2^2}. \quad (8)$$

$\nabla I = (Ix, Iy)$ is the spatial power inclination. For culmination purposes, we incorporate a dialog of typical speed in our review. The Lucas and Kanade optical flow algorithm permits the calculation of typical speed yet the Horn and Schunck optical flow calculation does not. Note that the 2D movement imperative condition can be re-composed as:

$$\vec{v}_n \cdot \hat{n} = v_n \quad (9)$$

Which is equivalent to equation (6), since the unit direction of normal velocity is $\hat{n} = \frac{(Ix, Iy)}{\|(Ix, Iy)\|_2}$ and the magnitude of normal velocity is $v_n = \frac{-It}{\|(Ix, Iy)\|_2}$ as in equation (7)

According to Fleet and Weiss, a robust optical flow estimation can be obtained by choosing the gradient based approach which uses pixel intensity translation for the flow [24]. Figure 3 shows the gradient constraint relates the displacement of the signal to its temporal difference and spatial derivatives (slope). For a

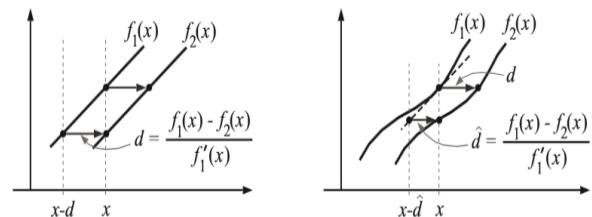


Figure 2 Gradient constraints on Linear and Non-Linear signals

displacement of a linear signal (left), the difference in signal values at a point divided by the slope gives the displacement. For

nonlinear signals (right), the difference divided by the slope gives an approximation to the displacement.

It was mentioned in their work that a common starting point for optical flow estimation was to assume that pixel intensities are translated from one frame (figure 3) to the next. Based on Horn and Schunck's estimation, the intensity was derived as follows:

$$I(\vec{x}, t) = I(\vec{x} + \vec{u}, t + 1) \quad (1.1)$$

Where $I(\vec{x}, t)$ is image intensity as a function of space $\vec{x} = (x, y)^T$ and time t , and $\vec{u} = (u_1, u_2)^T$ in the 2D velocity. To derive an estimator for 2D velocity \vec{u} , 1D case was considered first. Let $f_1(x)$ and then $f_2(x)$ assigned 1D signals (images) at two time instants. As shown in Figure 3, assuming that $f_2(x)$ be a interpreted version of $f_1(x)$; i.e., let $f_2(x)=f_1(x-d)$ where d means translation. A Taylor series expansion of $f_1(x-d)$ about x is given by

$$f_1(x-d)=f_1(x)-df_1'(x)+O(d^2f_1''). \quad (1.2)$$

where $f' \equiv df(x)/dx$. This expansion can be rewritten as the difference between the two signals at location x as $f_1(x)-f_2(x)=df_1'(x)+O(d^2f_1'')$. Ignoring second- and higher-order terms, An approximation can be obtained to d:

$$d = \frac{f_1(x)-f_2(x)}{f_1'(x)}. \quad (1.3)$$

The 1D case simplifies to 2D. As above, assuming the displaced image is approximated very well by a first-order Taylor series:

$$I(\vec{x}, +\vec{u}, t + 1) \approx I(\vec{x}, t) + \vec{u} \cdot \nabla I(\vec{x}, t) + I_t(\vec{x}, t) \quad (1.4)$$

Where $\nabla I \equiv (Ix, Iy)$ and I_t denotes spatial and temporal partial derivatives of the image I , and $\vec{u} = (u_1, u_2)^T$ denotes the 2D velocity. The following were the assumptions Lukas-Kanade made to accomplish their goal:

- Distant light source: the distance between light source and the object in the scene should remain constant
- No rotation of objects in the scene
- No secondary illumination (i.e. no shadows, no reflections of surfaces)

To achieve their goal, [24] used iterative coarse-to-fine refinement, different forms of parametric motion models, different conservation assumptions, probabilistic formulations, and robust mixture models were combined.

B. Optical flow application

Although optical flow has not seen the full flexed implementation in robot navigation, it has distinctly been applied in several areas of robotics. The applications of optical flow include time-to-collision estimation, pattern recognition, movement detection and tracking and visual odometry. The most prominent applications of it include pattern recognition, time-to-collision estimation, movement detection and visual odometry.

• Pattern Recognition

This area focuses on aspect of image processing including facial recognition which can be achieved through computed results of optical flow. Pattern recognition is one of the vital

and helpful branch in robotics today. It can make stock taking in large supermarkets very easy.

• Time-to-collision

Time-to-collision is also referred to as time to contact. Processing an image to obtain motion vectors can be used to calculate the estimated time to contact [26]. The optical flow object may contain information such as the magnitude of component of motion along a radial line from the focus of expansion which can be used to characterize the separation to the position of the vector concerning the focus of expansion (FoE) as p , and the part of movement along the spiral line as v , and time to contact as T , at that point then $T = \frac{p}{v}$. Hence the computed result from optical flow can be sampled and processed further to obtain information that can be used to estimate time to contact on autonomous robots [26]. It has been affirmed that time-to-collision (T_c) information can be obtained from optical flow based solely on the processing of target expansion rate (optic variable τ) [27].

• Movement detection and tracking

Detection and tracking have gained a lot of interest in the last few years [28]. Many systems use motion detection objects for tracking, detecting by locating blobs of motion and computing the vertical histogram of the outline. Such system from the researchers points of view can be used effectively to segment groups of people [29]. In order to track an object, it must be able to reliably and consistently detect and also have features that can be observed and matched from frame to frame and be able to extract. Motion estimation and following are enter exercises in numerous PC vision applications, including action acknowledgment, activity checking, car wellbeing, and reconnaissance [30] which can be achieved by using optical flow object.

• Visual Odometry

Visual odometry involves the activities used in determining the position and orientation of a robot by analyzing the associated camera images [31]. Visual odometry algorithms are more fruitful in open air than indoor territories as it is less demanding to extricate more elements contrasted with highlights removed in indoor conditions. Case of visual odometry is the work of Nister [32].

V. REASONS FOR INFREQUENT HOLISTIC APPLICATION OF OPTICAL FLOW IN ROBOT NAVIGATION

In an attempt to find out the reasons behind the rareness of the holistic application of optical flow for robot navigation, it was realized that, no specific materials or publication has categorically stated the actual cause of the infrequent holistic

application of optical flow in robot navigations but the comments made by some of the reviewed works geared toward the cost associated with the computation of the optical flow itself. For instance, Cui, et al trying to improve on existing optical flow algorithm said the drawback of optical flow approach was the large computational cost [33]. Also, the survey made by Denis et al. [34] on the various optical flow computational methods had the cost of computation being one of the prevailing limitations of such methods.

VI. RELATED WORK

For a few years, there has been enthusiasm for the utilization of optical stream for vision based versatile robot route, with the mean to accomplish energetic execution for navigational assignments. This is conceivable because of an observation that optical flow methods are computationally costly and do not have the precision and strength required for utilized as a part of the control circle. Once more, the essential worry for optical stream is the productivity and the shabby estimate as the stream field is regularly gotten. Such approximations regularly require post preparing and accepted condition structure so as to acquire a workable control input. For an entire robot framework that depends on optical stream for numerous parts of control, the utilization of full optical stream estimation methods is alluring. The decision however stays as a challenge.

Selim Termizer developed algorithm for computing optical flow which based on the theory of edge detection for optical flow computation. In his work, images were continually requested by the navigation system to serve as the basis for computation of the optical flow. Noise effect on read images were reduced by using low pass filters and edges were found by applying Laplacian filter to the already filtered images before patch matching were made to find the Optical flow [35]. In this research, similar optical flow algorithm were used for the computation but the computational system concept differ: the image processing can be dissociated from the robot navigation system, also, a fundamental OF algorithm was considered to prove efficiency. In the robot, thresholds are received and based on for navigations. In Robotics inquire about, the utilization of visual contribution for route purposes began in late 1970's. Among the primary employments of cameras for portable robot route can be followed to Moravec's work, who used several cameras to navigate a robotic cart in a room [36]. However, the use of vision in mobile robotics has been hindered by the limited computational power. Image processing and understanding tasks require much computational power due to the amount of data in images, which was not available until recent advances in hardware computer vision algorithms.

VII. METHODS

After assessing the design of many similar works, a model of the system which was used to perform navigations based on optical flow was created as seen in figure 4. The entire system comprised of three major subsystems: the robot, optical flow

computation system, and the navigation system. The latter two constituted the schema of the robot system in reality.

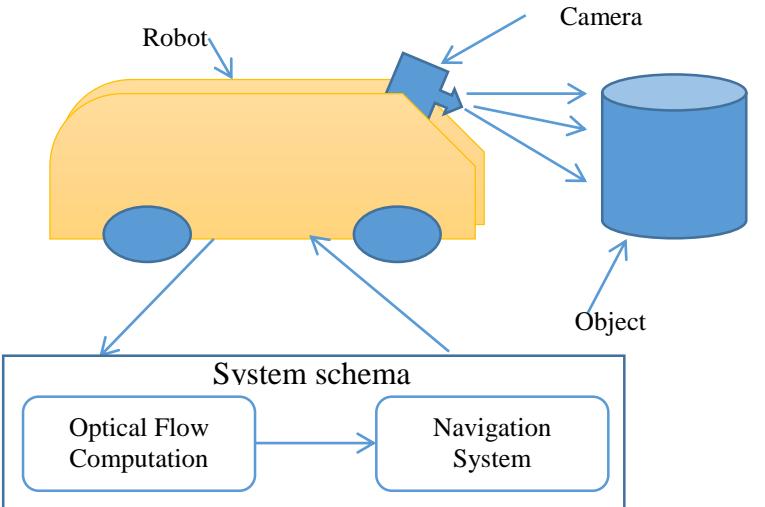


Figure 4 *Optical Flow navigation System*

The robot component comprised of all the actuators and sensors for its motion. For the purpose of experiments, a virtual robot was designed using Matlab's Simulink. During the experiment a webcam with frame dimension 320 X 240 served as the robot's eye. In the real world implementation, the camera and the images' dimension can be set to any required value, but in the research, an image dimension of 640 X 480 and frame rate of 50fps were used for the image processing. It should be noted that, the chosen dimension would affect the threshold for the navigation. A box shape geometry object in the virtual world represented the robot in the virtual environment. The OF computational subsystem comprised of Lucas-Kanade (LK) algorithm implementation for estimating the optical flow. This subsystem received its input as a sequence of captured images or video stream from the robot eye locally; but the intended design could also receive images transferred remotely. The format of the realtime video used was avi file. The results from the computation subsystem were passed on to the navigation subsystem which then performed navigation based on the sampled components of the optical flow resulted value.

A. Designed optical flow navigation Algorithm

Step 1: Read optical flow value from the computation sub-system output

Step 2: When concentration does not match the center threshold (that is when there is less flow at center), Go Straight forward. Otherwise,

Step 3: When Optical flow concentration is more at right of the frame of reference than it is at the left and meets the left threshold, Turn Left. Otherwise,

Step 4: When the flow concentration is more at left than at the right and meets right threshold Turn Right. Otherwise,

Step 5: When the flow concentration of the front eye is more at left, right and center but the back eye meets the left condition such as of the front eye but for right, Reverse Right. Otherwise,

Step 6: when the flow concentration of the front eye is more at left, right and center but the back eye meets the left condition such as of the front eye but for right, Reverse Left. Otherwise, **Step 7:** when the flow is more in all directions ahead but the back eye has less concentration at the center than that of the left and right, Reverse Straight. Otherwise,

Step 8: When the flow concentration of the front eye is more at left, right and center and same condition is met at the back eye go to step 9 otherwise go to step 1.

Step 9: If robot eyes still sees, go to step 1, otherwise stop.

There is less or more flow at a region if vertical or horizontal component values of the computed optical flow result are less or more respectively. The reverse navigations shall be performed if and only if back is not blocked through a rear eye of the robot. Since the forward and reverse algorithms are same, only the forward navigation was experimented. The algorithm for the navigation was implemented in Matlab for the experiments.

B. Tests and Result

Figure 5 shows the design of the environment that was expected to prove optical flow based robot navigations. After a number of

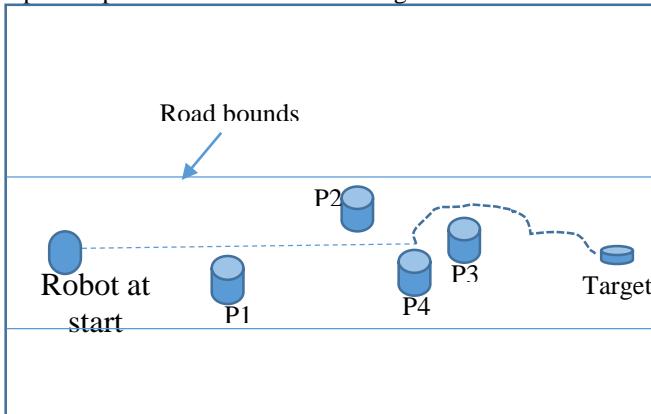


Figure 5 Navigation Environment

tests and analysis, a thresholds were obtained for the left, right and forward navigations out of the computed optical flow object. In the environment the p1 to p4 were objects which moved in their respective positions that caused the robot to react in a certain way as and when the robot saw them. The robot was able to perform navigations avoiding obstacles and also reaching its target goal. The LK algorithm was understudied and used for the optical flow computation to provide efficient navigation for the robot in a virtual world.

VIII. CONCLUSION

The face of AI would completely change together with world's technology at large when fully flexed efficient robot navigation is realized. Applications. Although Krajiník [37] stated that robot navigation is a complex, technological problem as it determines a robot's autonomy and reliability in performing assigned tasks and that it has been widely researched since the 1970s. Again, many solutions and techniques have been proposed, the navigation problem remains challenging. Many of the proposed solutions were blind type of navigation algorithm

– the navigations were performed just like a blind folded person would navigate through an environment. Works reviewed show that, it has been widely accepted that, humans and other animals base on optical flow for their navigations. Since the initiation of the concept of Optical Flow by Gibson (1950), several researchers developed ideas for its computation which in turn has been improved over the last decades. This optical flow has been applied distinctly in most robotic fields such as time-to-collision estimation, pattern recognition, movement detection and tracking and visual odometry.

Technology advancement in our current world depicts that future would not be able to do away with robots looking at the trend of technology developments. Industries and even homes are making use of robots being it stationary or mobile. But non-vision components (such as sensors such as sonars, radars, etc.) are mostly still in use for autonomous robots navigations which are less informative than optical flow. With optical flow several parameters can be obtained for the navigation especially for the purposes of detecting objects to avoid objects.

The experiments conducted in this research clearly reveals that, autonomous robots can easily perform their necessary navigations using cheaper vision component such as cameras, based on optical flow. The base OF algorithm such as Lukas-Kanade was selected on the basis of confirming efficiency. It is believed that, LK algorithm has been the base algorithm which means it's efficiency is low as compared to many of the newer algorithms which are improved version of LK. After several experimentation and adjustment of thresholds, the agents was able to navigate based on the calculated OF values from the LK algorithm.

Although all the experiments were conducted in a simulation environment, it cannot be overemphasized that the selected simulation environment (i.e. Matlab's Simulink) has performed tremendous positive implementations through its simulations for many industries and science bodies such as NASA as said in MATWORKS [38] bases on its experiments before actual production. Although it is easy to argue against the calculation intensive nature of the optical flow algorithm, it cannot be overemphasized that the exponential growth nature of technology is producing processors that are powerful enough to implement the optical flow computations with ease and hence robot navigation can efficiently be performed using optical flow. Reviews during this research revealed that, optical flow can be applied in several distinct aspects of robotics. It should be noted that, optical flow has the greatest potential to be applied for robot navigations rather than the other expensive methods such as Radar, Sonar and Lidar and should therefore be embraced and continually developed for robot navigation purposes that may combine two or more of the distinct applications. Consider a robot seeing you and able to mention your name, shakes hands with you and move pass by you without bumping into you or any other object: the goal of computer vision would be achieved completely in that light then. Future work should consider object identification during navigations for more precise robot decision making on navigation

REFERENCES

- [1] J. Ng, "An Analysis of Mobile Robot Navigation Algorithms in Unknown Environment," School of Electrical, Electronics and Computer Engineering, UWA, 2010.
- [2] J. K. Worrall, "Guidance and Search Algorithms for Mobile Robots: Application and Analysis within the Context of Urban Search and Rescue," Department of Electronics and Electrical Engineering, University of Glasgow, Glasgow, 2008.
- [3] F. Glover, "Tabu Search: Part 1," *ORSA J. of Computing*, vol. II, no. 1, pp. 190-206, 1989.
- [4] E. McGookin and D. Murray-Smith, "Submarine Manoeuvring Controllers' Optimisation Using Simulated Annealing and Genetic Algorithms," *Control Engineering Practice*, vol. I, no. 14, pp. 1-15, 2006.
- [5] R. Montemanni, L. Gambardella and A. Das, "The minimum power broadcast problem in wireless networks: a simulated annealing approach," *IEEE Wireless Communications and Networking Conference*, vol. 4, no. 15, pp. 2057-2062, 2005.
- [6] C. Ellis, "A Bluffers Guide to Genetic Algorithms," *Engineering Design Newsletter*, SERC, Summer, 1993.
- [7] J. Tompkin, "Optical Flow An Introduction," *Machine Vision - Practical 2*, pp. 1-13, 13 March 2008.
- [8] D. C. McCarthy, "Performance of Optical Flow Techniques for Mobile Robot Navigation," Department of Computer Science and Software Engineering, The University of Melbourne, Parkville, Victoria, Australia, 2005.
- [9] J. J. Gibson, "The Perception of the Visual World," *Audio Visual Communication Review*, vol. Vol. 1, no. 3, pp. 190-194, 1953.
- [10] D. Kondermann et al, "On performance Analysis of Optical Flow Algorithms," Interdisciplinary Center for Scienti, Heidelberg, 2012.
- [11] M. Drulea and S. Nedevschi, "Total variation regularization of local-global optical flow," *Computer Vison*, pp. 1-6, 13 May 2011.
- [12] Z. Wei, D.-J. Lee and B. E. Nelson, "FPGA-based Real-Time Optical Flow Algorithm Design and Implementation," *Journal of Multimedia*, vol. 2, no. 5, pp. 28-45, 2007.
- [13] M. Ye, "Robust Visual Motion Analysis: Piecewise-Smooth Optical Flow and Motion-Based Detection and Tracking," University of Washington, Washington, 2002.
- [14] V. S. Chakravarti, Y. J. M. Shirur and P. Rekha, "Communication, Advanced Devices, Signals Systems and Newtorking," VCASAN, 2013.
- [15] H. Foroosh, J. Zerubia and M. Berthod, "Extension of Phase Correlation to Subpixel Registration," *IEEE Transaction on Image Processing*, vol. 11, no. 3, pp. 188-200, 2002.
- [16] C.-L. Liu, C. Zhang and L. Wang, "Pattern Recognition," Beijing, China, 2012.
- [17] H. S. Stone, "A Fast Direct Fourier-Based Algorithm for Subpixel Registration of Images," *IEEE Transactions On Geoscience and Remote Sensing*, vol. 39, no. 10, pp. 2235-2243, 2001.
- [18] B. Atcheson, W. Heidrich and I. Ihrke, "An Evaluation of optical flow algorithm for background oriented Schlieren Image," Department of Computer Science, University of British Columbia, Canada, 2008.
- [19] G. W. Humphreys and V. Bruce, "Visual Cognition," Psychology Press, ISBN 0-86377-124-6, 1989.
- [20] S. S. Beauchemin and J. L. Barron, "The Computation of Optical Flow," ACM New York, USA, 1995.
- [21] A. Bruhn, J. Weickert and C. Schnörr, "Combining the Advantages of Local and Global Optic Flow Methods," Berlin Heidelberg, 2002.
- [22] B. Glocker, N. Komodakis, G. Tziritas, N. Navab and N. Paragios, "Dense Image Registration through MRFs and Efficient Linear Programming," *PubMed*, pp. 1-2, 7 April 2008.
- [23] C. Lei and Y. Yee-Hong , "Optical Flow Estimation on Coarse-to-Fine Region-Trees using Discrete," *Computer Vision, IEEE 12th International Conference*, vol. XII, pp. 1562-1569, 2009.
- [24] D. J. Fleet and Y. Weiss, "Optical Flow Estimation," *Mathematical Models in Computer Vision*, vol. 15, pp. 239-258, 2005.
- [25] J. L. Barron and N. A. Thacker, "Tutorial Computing 2D and 3D Optical Flow," *Features and Measurement Series*, pp. 1-5, 20 January 2005.
- [26] OPTICAL FLOW, "<http://opticflow.bu.edu/research/time-to-contact-estimation/>," 2011. [Online]. Available: <http://opticflow.bu.edu>. [Accessed 20 October 2012].
- [27] M. Laurent, "Visual Information and Skill Level in Time-To-Collision Estimation," Arcueil Cedex, France, 1985.
- [28] M. F. Abdelkader, R. Chellappa , L. A. Chan and A. L. Chan, "Integrated Motion Detection And Tracking for Visual Survellence," *Fourth IEEE International Conference on Computer Vision Systems*, p. 28, 7 January 2006.
- [29] I. Haritaoglu, D. Harwood and L. Davis, "A real time system for detecting and tracking in people 2 1/2D," in *5th European Conference Computer Vision*, 1998.
- [30] D. Matlab R2016a, "Object Tracking and Motion Estimation," 2016. [Online]. Available: <http://www.mathworks.com/help/vision/object-tracking-and-motion-estimation.html>. [Accessed 30 March 2016].
- [31] M. Maimone, L. Matthies and Y. Cheng, "Two years of Visual Odometry on the Mars Exploration Rovers," *Journal of Field Robotics*, vol. 24, no. 3, p. 169–186, 2007.

- [32] D. Nister, O. Naroditsky and J. Bergen, "Visual Odometry," in *International Conference Computer Vision and Pattern Recognition*, 2004.
- [33] G. Cui, X. Chen and J. Guo, "Visual Robot Navigation based on Improved Optical Flow Algorithm and Optimized Bessel Curves," *Information Technology*, vol. 13, pp. 2363-2368, 2014.
- [34] F. Denis, P. Boutry and C. Kerfrann, "Optical Flow Modeling and Computation: A Survey," *Computer Vision and Image Understanding*, pp. 1-21, 26 February 2015.
- [35] T. Selim, "Optical Flow for Robot Navigation," *Artificial Intelligence MIT*, pp. 1-3, 2010.
- [36] H. Moravec, "Towards automatic visual obstacle avoidance," in *5th International Joint Conference Artificial Intelligence*, 1997.
- [37] T. Krajiník, V. Vonásek, D. Fišer and J. Faigl, "AR-Drone as a Robotic," Heidelberg, 2011.
- [38] MATWORKS,
] "<https://www.mathworks.com/matlabcentral/answers/46502-does-matlab-application-using-image-processing-is-used-in-nasa>," 2012. [Online]. Available: <https://www.mathworks.com>. [Accessed 23 May 2013].
- [39] J. Barron, D. Fleet, S. Beauchemin and T. Burkitt, "Performance of optical flow techniques," *CVPR*, pp. 23-27, 1992.
- [40] A. Burton and J. Radford, "Thinking in perspective: critical essays in the study of thought processes," London, 1978.

AUTHORS PROFILE

Nixon Adu-Boahen holds an MPhil. Computer Science from the Kwame Nkrumah University of Science and Technology (KNUST). He had his BSc. Computer Science degree KNUST. He served at the Department of Computer Science as a National Service Person where he provided tutorials to students. He is currently working at Garden City University College as a Tutor. Nixon has keen interest in robotics and databases and analytics.

Dr J. B. Hayfron-Acquah received the BSc degree in Computer Science from the Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana, his MSc Computer Science and Applications degree from Shanghai University of Science and Technology, Shanghai, China and his PhD from the Southampton University, Southampton, England. He is currently a Senior Lecturer at the Department of Computer Science, KNUST. He has over 40 publications to his credit. His research areas include Biometrics, Cloud Computing, Networking, Image Processing and Computer Security.

J. K. Panford received his BSc degree in Computer Science from the Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana, his MSc Software Technology degree from Stuttgart University, Stuttgart, Germany. He is currently a Lecturer and pursuing his PhD in Computer Science at the Department of Computer Science, KNUST. He has over 15 publications to his credit. His research areas include Cloud Computing, Networking, Image Processing and Computer Security, Software technology and Embedded Systems.

Dataset of Graphs and Sub-graphs: Storage Representation and its Graphical form

Rachna Somkunwar¹ and Dr. Vinod M. Vaze²

¹ JJT University, Research scholar, Rajasthan, India, rachnasomkunwar12@gmail.com

² JJT University, Assistant Professor, Rajasthan, India, vinod.vaze@gmail.com

Abstract- Graphs are used to represent the structure of objects which has set of vertices and a set of edges. Graphs are powerful and universal data structure useful in various fields of computer science and engineering. Various tools and programs are available to generate graphs in binary format but it's very difficult to read binary Graphs. For matching graphs and sub-graphs, dataset are created first. In this paper dataset is created for different types of graphs and sub-graphs like Random Graphs, M2D, M3D, and M4D Graphs. This paper aims at to generate graphs in binary format and represent these binary graphs in an image form. For generating binary graphs, programs have been used and stored in the dataset. A number of experiments have been done for generating different binary graphs for creating the dataset.

Keywords: Graph, Sub-graph, Dataset, Binary Graphs, DOT file.

I. INTRODUCTION

Graphs are an effective representation for organizing data, i.e. data that can be effectively portrayed by its subparts and the relations between these subparts. The classification of input samples represented by graphs is not trivial, since the most commonly used classifiers are based on a vectorial representation. The comparison of two graphs can be performed using graph matching techniques, but they are in the general case very expensive. Other than the grouping, graphs can be utilized additionally for the issue of inductive learning, i.e. given an arrangement of illustrative graphs divided into classes, finding an appropriate depiction of the attributes of each class that can be utilized to group future examples.

Boolean algebra frames a foundation of software engineering and advanced framework plan. Numerous issues in advanced rationale outline and testing, computerized reasoning, and combinatorics can be communicated as an arrangement of operations on Boolean capacities [1]. Two Boolean expressions signify a similar capacity (equality) require answers for NP-Complete problems [2]. Consequently, all known approaches to performing these operations require, in the worst case, an amount of computer time that grows exponentially with the size of the problem.

In recent years a huge amount of algorithms for classification, clustering, and analysis of objects given in terms of feature vectors have been developed [3].

Due to the ability of graphs to represent properties of entities and binary relations at the same time, a growing interest in graph-based object representation in pattern analysis can be observed [4]. That is, graphs found widespread applications in science and engineering. In the fields of Bioinformatics and Chemo informatics, for instance, a graph based representations have been intensively used [5, 6, 7].

Another field of research where graphs have been studied with emerging interest is that of web content mining [8].

Graphs can be stored in two different ways in file format: Text format and Binary format. To store graphs in Text format, standard file format is used known as DOT. DOT is specially designed to store the information of the graph. In this paper, dataset of graphs and sub-graphs are created in binary form.

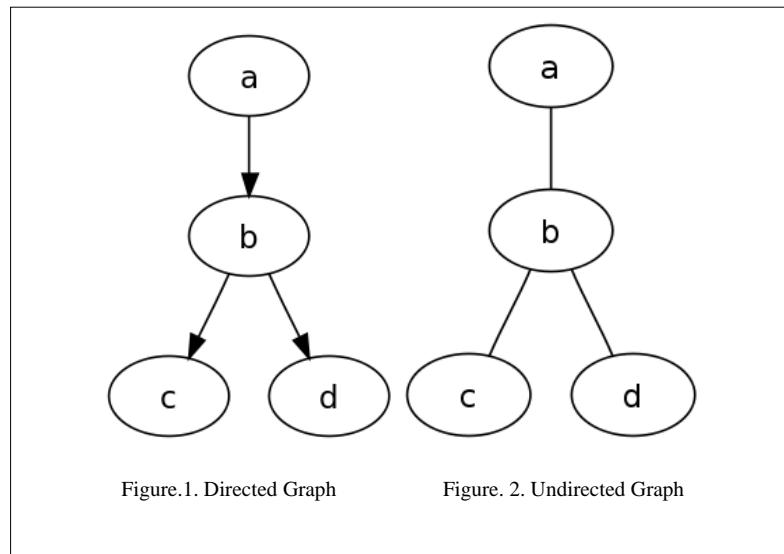
A. Basic Terms

DOT File: DOT is a graph description language, which is used to store the data in text form. For handling Big Graphs, it requires more space. Hence Dot file is best suitable for small graphs. Extensions of such kinds of files are denoted by DOT. Directed and Undirected graphs are described by DOT. C and C++ languages are supported by DOT.

Graphs are classified as Directed, Undirected, Labeled and Unlabeled Graphs.

Directed Graphs: Directed Graphs are also called as Diagraph. If edge e is having direction in the graph, then edge e is said to be directed edge. For directed graph order of vertices matters a lot. Figure 1 shows Directed Graph, where the arrow shows the relationship between nodes.

Undirected Graphs: Undirected graphs are those graphs in which directions on the edges are not given, shown in figure 2. This type of graphs represents relationships between objects.



Degree of Vertex: Total number of incoming and outgoing edges of a vertex is called as Degree of a vertex.

Labeled Graphs: Sometimes labeled are not necessary if we are dealing with unlabeled graphs. A graph is said to be unlabeled if vertices and edges are not having any value, otherwise it comes under the labeled graph category [9].

Random graphs are drawn at random where vertices are added with n number of disconnected vertices. Random graphs are shown in figure 3. Randomly connected graphs are graphs whose density value is different.

Bounded valance graphs are those graphs where a limit is specified. Every vertex may have a number of links and that should be less than the specified limit, it is said to be valance.

Fixed valance graph, where the number of vertices is equal to the number of links.

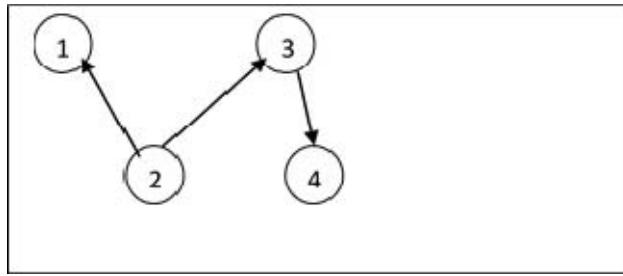


Figure. 3. Random Graph

Regular Mesh: Vertices degree will be same. It is represented by different size like 2D, 3D, 4D. 2D mesh means square of N number of nodes, 3D is a cube of N number of nodes and so on.

Irregular Meshes: It is a combination of regular meshes and random connected edges with equal allocation [10].

Planar graphs are represented in planes where edges do not cross to each other [11].

B. Graph Representation in Memory

The main idea is how much memory or space is required to store a graph in memory. There are two ways to represent a graph in memory, first is Sequential and second is Linked list.

Sequential: Storing a graph in matrix form is sequential. It is very easy ways to implement as operations are represented by a matrix form [12]. It is faster method, but in case of huge graphs it is not applicable.

if $a_{ij} = 1$ vertex i is neighbor of vertex j.

else $a_{ij} = 0$.

For undirected graphs, the adjacency matrix is symmetric. Symmetric means only diagonal elements need to be stored and for undirected graph it is not symmetric. The advantage of adjacency matrix is we can add and search edges which require a linear amount of memory.

Linked List is Storing a graph in the linked list is Linked list representation. Instead of using adjacency matrix for huge graph we can use the adjacency list which requires less memory.

Adjacency matrix is good for Dense Graph. Adjacency List is good for Sparse Matrix.

1D array represents vertices and 2D array represent edges.

V stands vertices and E stands edges. K is the number of vertices adjacent to vertex.

Memory Complexity for adjacency matrix is $O(|V|^2)$ and for adjacency list is $O(|E|)$. Adding a new edge may require $O(1)$ for adjacency matrix and Linked List. Removing an edge may require $O(1)$ for adjacency matrix, but for linked list it may require $O(K)$. Similarly for searching an edge may require $O(1)$ for adjacency matrix and $O(K)$ for the linked list.

II. DATASET

In this paper dataset is created for different types of graph and sub-graph. Different types of graph are randomly generated graphs, M2D, M3D, M4D graphs. For any algorithm, it is mandatory to have different sizes of graphs.

Table 1. Dataset for Graphs

Types of graph	Number of Nodes	Number of test-set graph	Number of data-set graph
General/M2D/M3D/M4D	100	1000	1000
	200	1000	1000
	300	1000	1000
	400	1000	1000
	500	1000	1000
	600	1000	1000
	700	1000	1000
	800	1000	1000
	900	1000	1000
	1000	1000	1000
	2000	1000	1000
	3000	1000	1000
	4000	1000	1000
	5000	1000	1000
	6000	1000	1000
	7000	1000	1000
	8000	1000	1000
	9000	1000	1000
	10000	1000	1000

Table 2. Dataset for Sub-graphs

Types of sub-graph	Maximum number of Nodes	Number of test-set sub-graph (30% of full graph)	Number of data-set of full-graph
General/M2D/M3D/M4D	100	1000	1000
	200	1000	1000
	300	1000	1000
	400	1000	1000
	500	1000	1000
	600	1000	1000
	700	1000	1000
	800	1000	1000
	900	1000	1000
	1000	1000	1000
	2000	1000	1000

Table1 and Table2 show dataset of graphs and sub-graphs. It contains Number of nodes, number of test-set graph and number of data-set graph. While creating the dataset it is necessary to have graphs with different sizes like 100,200,300,400,500,600,700,800, 900, 1000. These types of large graphs are suitable for Graph and Sub-graph Isomorphism.

III. PROPOSED METHOD

To check Graphical representation of graphs, a small program is created. It is shown in figure 4.

Input: Binary Graph

Output: Graph Image

1. Graph reader reads the input binary file and generates its appropriate adjacency matrix form.
2. Dot Converter converts the adjacency matrix to DOT file.
3. DOT file is converted into its png form to get its Graphical Representation.

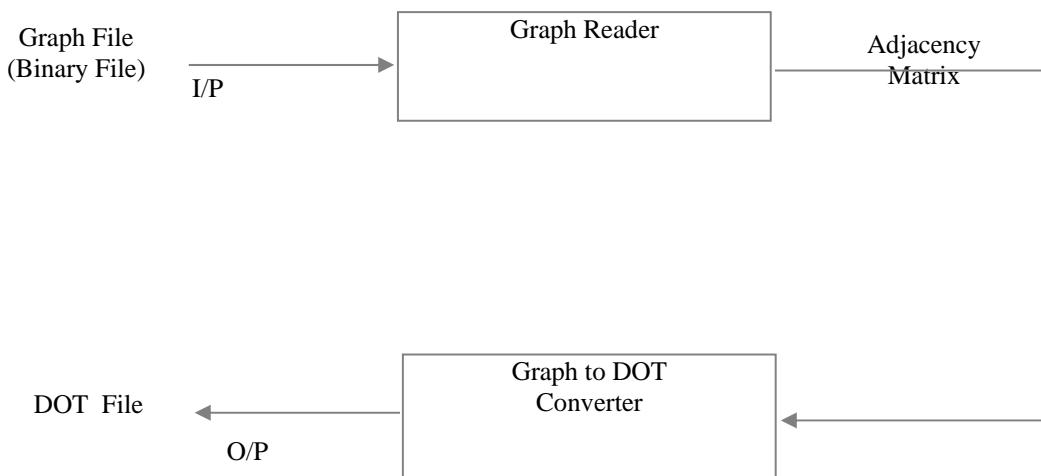


Figure 4. Block diagram to convert Binary file into DOT file

After conversion of Binary file to DOT file, DOT file is converted to its graphical representation. DOT is used to create images in jpg/png form, which contains information of graph. Following command is used to convert DOT file to Image file.

```
$> dot -Tpng graph1.dot -o graph1.png
```

Graph1.dot DOT file and graph1.png is png form of graph1.DOT.

IV. RESULT

For the generation of different types of graphs, Graph Generator program is implemented to generate the data set of Graphs and Sub-graphs.

For the generation of graphs, different types of computer programs of 'C' language are developed by <http://www.mivia.unisa.it> for the generation of different types of graphs.

```
$>/gen.out <no of nodes> <no of edges> <dataset file> <testset file>
```

Example:

\$>/gen.out 10 20 file1 file2

gen.out requires 4 arguments
Argument 1: no of nodes =10
Argument 2: no of edges =20
Argument 3: dataset file = file1
Argument 4: testset file = file 2

Node Count=10 (gen 10 nodes and 20 edges)

```
1 0 0 1 0 0 0 0 0 0  
1 1 1 0 0 0 0 0 0 0  
0 1 1 1 0 1 1 0 0 0  
1 0 0 1 0 0 0 0 0 0  
0 0 1 0 1 0 0 0 0 0  
0 1 0 1 0 1 1 0 0 0  
0 1 0 1 0 0 1 0 1 0  
0 1 0 0 0 1 0 1 0 0  
0 0 0 0 0 1 0 0 1 0  
0 0 1 1 0 0 0 0 0 1
```

```
digraph d  
{  
v1[label=1]  
v2[label=1]  
v3[label=1]  
v4[label=1]  
v5[label=1]  
v6[label=1]  
v7[label=1]  
v8[label=1]  
v9[label=1]  
v10[label=1]  
v1->v4[label=1]  
v2->v1[label=1]  
v2->v3[label=1]  
v3->v2[label=1]  
v3->v4[label=1]  
v3->v6[label=1]  
v3->v7[label=1]  
v4->v1[label=1]  
v5->v3[label=1]
```

```
v6->v2[label=1]
v6->v4[label=1]
v6->v7[label=1]
v7->v2[label=1]
v7->v4[label=1]
v7->v9[label=1]
v8->v2[label=1]
v8->v6[label=1]
v9->v6[label=1]
v10->v3[label=1]
v10->v4[label=1]
}
```

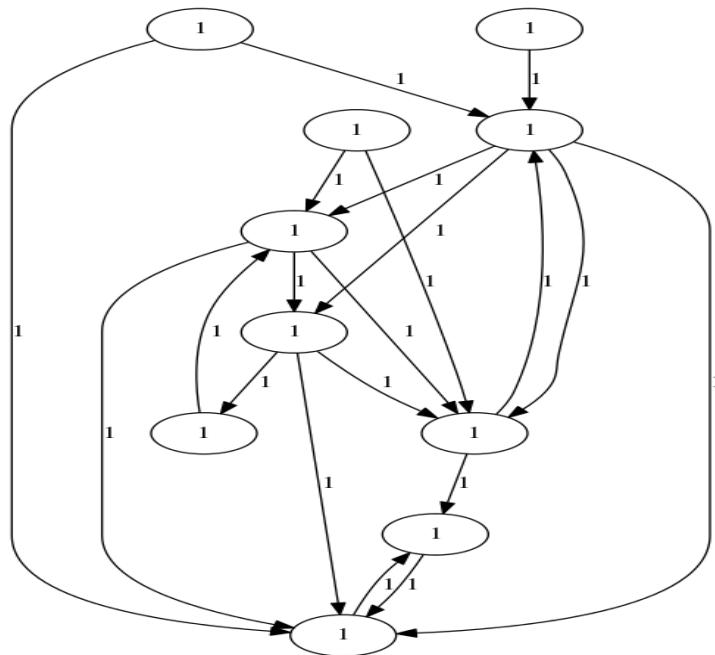


Figure 5. Graphical representation of Gen graphs with 10 nodes and 20 edges

V. CONCLUSION

In this paper a new method is discussed for reading a binary graph and generates its image form. It is based on the idea of first creating a dataset of graphs. A dataset of graphs is created for graph and sub-graph in binary form. The experiments confirmed that image form is created using -Tpng command. We conclude that the proposed method presented in this paper is highly recommended for dataset creation and generating its graphical form. The dataset is created without knowing whether it is in the correct form or not, to verify dataset of graphs, DOT file is converted

into its graphical form. Numerous experiments have been done for creating dataset and verified by the proposed method.

REFERENCES

- [1] Bryant, Randal E. "Graph-based algorithms for boolean function manipulation." *Computers, IEEE Transactions on* 100.8 (1986): 677-691.
- [2] M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*.
- [3] Duda, R., Hart, P., Stork, D.: *Pattern Classification*, 2nd edn. Wiley Interscience, Hoboken (2000)
- [4] Conte, D., Foggia, P., Sansone, C., Vento, M.: Thirty years of graph matching in pattern recognition. *Int. Journal of Pattern Recognition and Artificial Intelligence* 18(3), 265–298 (2004)
- [5] Borgwardt, K., Ong, C., Schönauer, S., Vishwanathan, S., Smola, A., Kriegel, H.P.: Protein function prediction via graph kernels. *Bioinformatics* 21(1), 47–56 (2005)
- [6] Mahé, P., Ueda, N., Akutsu, T.: Graph kernels for molecular structures – activity relationship analysis with support vector machines. *Journal of Chemical Information and Modeling* 45(4), 939–951 (2005)
- [7] Ralaivola, L., Swamidass, S., Saigo, H., Baldi, P.: Graph kernels for chemical informatics. *Neural Networks* 18(8), 1093–1110 (2005)
- [8] Schenker, A., Bunke, H., Last, M., Kandel, A.: *Graph-Theoretic Techniques for Web Content Mining*. World Scientific, Singapore (2005)
- [9] Joseph A. Gallian, “A Dynamic Survey of Graph Labeling”, *the electronic journal of combinatorics*, 2011.
- [10] P.Foggia, C.Sansone, M.Vento, “A Database of Graphs for Isomorphism and Sub-graph Benchmarking”, University of Naples, “Federico II”.
- [11] H. De Fraysseix, J.Pach and R. Pollack, “How to Draw a Planar Graph on a Grid”, *Combinatorica*, 1990.
- [12] Alan Curtiss Tucker, “Matrix characterizations of circular-arc graphs”, *Pacific Journal of Mathematics*, vol. 39, 1971.

Fingerprint Image Retrieval using Statistical Methods

Sudhir Vegad

Department of Information Technology
A D Patel Institute of Technology
New V V Nagar, India
svegad@gmail.com

Dr. Tanmay Pawar

Electronics Department
BVM Engineering College
Vallabh Vidyanagar, India
tanmaypawar@gmail.com

Abstract— The image is composed of information and pixels. The Information of image is in brightness of Red, Green and Blue Channel are used for hiding data. Image retrieval problem encountered when finding and retrieving images that is similar to a user's query from a database. Done content built picture retrieval, enter dives in the manifestation from claiming a picture. Over these images, distinctive offers are concentrated on the opposite pictures starting with database would retrieved appropriately. Biometrics recognizes the individuals toward their physical or behavioral qualities. Fingerprints need aid seen as a champion around the large portion robust for mankind's distinction due to their uniqueness also creativity should recover finger impression pictures on the support about their different features. The fingerprints are taken from FVC2000, FVC2002 and FVC2004 and retrieved utilizing distinctive between GLCM, SURF and Gabor Wavelet. From those information finger impression image, as a matter of first importance focal point side of the point region is chosen as its textural characteristics would concentrated and saved clinched alongside different database.

Keywords- Image Retrieval, Fingerprint, GLCM, SURF, GBW

I. INTRODUCTION

The content-based picture recovery appears to be to need originated in 1992. The point when it might have been utilized by t. Kato on portray trials under programmed recovery about pictures starting with An database, In light of those shades What's more shapes available. Since then, the haul need been used to describe those transform for retrieving fancied pictures from an extensive accumulation on the groundwork for linguistic picture Characteristics. Those techniques, tools, and calculations that need aid utilized begin from fields, for example, statistics, design recognition, sign processing, and workstation dream. The soonest business CBIR framework might have been created Eventually Tom's perusing IBM Also might have been known as QBIC (Query by picture Content). Later system Furthermore chart based methodologies have exhibited a straightforward Also engaging elective should existing systems. Those enthusiasm toward CBIR needs to be developed due to the confinements inalienable to metadata-based systems, and additionally those huge go about workable employments to effective picture recovery. Printed majority of the data over pictures might be effectively searched utilizing existing technology, yet all obliges people with manually depict every picture in the database. This camwood a chance

to be illogical for altogether vast databases alternately to pictures that is formed repeatedly, the individuals after reconnaissance image capturing devices. That is similarly workable to error pictures that utilization diverse replacements clinched alongside their portrayals.

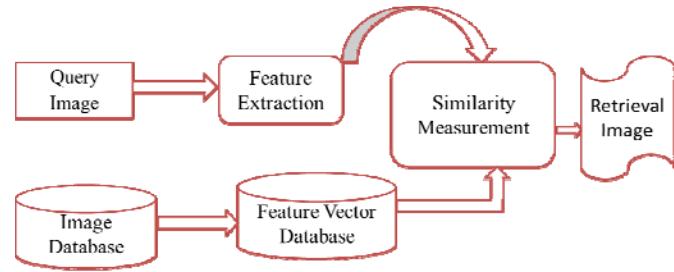


Figure 1: Basic CBIR Flow

It is depend on matching strategy. Image Retrieval Process is separated into two groups: Context Image Retrieval and Content Image Retrieval. Context suggestions the maximum accurate Info when images to be annotated with their appropriate names. A disadvantage of ABIR is manual picture explanations are tedious. Human comment is Subjective; a few Images couldn't be explained on the grounds that it is hard to depict their Content with Words. Shading highlight incorporate GLCM, SURF and BoG texture Features. For Fingerprint Feature Texture gives good results than Colour and Shape. Texture feature are very suitable and gives robust output because it's thoughtful of translation, rotation, and different scales patterns.

II. LITERATURE REVIEW

We have surveyed various techniques of feature extraction and image retrieval techniques.

In [1] the issue of finger impression order and also distinctive workable results were portrayed. Two systems were explained from claiming finger impression arrangement. Furthermore, rule-based indexing approach offers concentrated starting with the directional images and the variance-based neural system which offers work extension for accurate and faster systems.

In [2] Biometric-based frameworks also need a few limits that might have unfriendly suggestions to that security of a framework at the same time. A portion of the impediments of biometrics might make succeed with the advancement for

biometric engineering organization and a cautious framework design. It will be paramount to see the idiot proof personal distinguishment. Frameworks essentially don't exist and perhaps never will. Security may be a Hazard administration methodology that identifies, controls, Eliminates, alternately minimizes dubious occasions that might adversely influence framework assets. Furthermore majority of the data holdings those security level of a framework relies on the prerequisites (threat model) about a. requisition and the expense invade. In our opinion, legitimately executed biometric frameworks are successful deterrents to culprits.

III. DIFFERENT METHODS OF FINGERPRINT IMAGE RETRIEVAL

A. GLCM

There need of three sorts from claiming visual cues people regularly search for an image: ghastly (average tonal dialect variety, previously different groups for noticeable wavelengths), relevant (macro information surveyed from encompassing data), and textural. Textural information, or the spatial conveyance from claiming tonal dialect variety inside a band, is a standout amongst the greater part imperative aspects utilized within identikit Questions alternately districts from claiming enthusiasm toward a picture. Haralick, Shanmugam, Furthermore Dinstein presented a set about 13 composition features ascertained from an image's grey-level co-event grid (GLCM). These Haralick features, which would still generally utilized today for a extent for applications, permit quantification of a composition.

The GLCM calculates how frequently a pixel for gray-level (grayscale intensity) quality i happens whichever horizontally, vertically, or diagonally to contiguous pixels with the quality j .

GLCM direction of analysis:

- Horizontal
- Vertical
- Diagonal

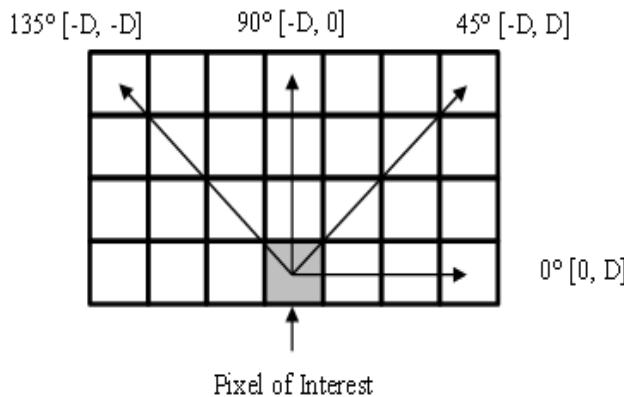


Figure 2: Directional Analysis for GLCM (P_0 , P_{45} , P_{90} & P_{135})

B. SURF

Surf offers will a chance to be a scale-invariant trademark distinguish in perspective from claiming the individuals hessian matrix, similarly is, those Hessian-Laplace distinguish [21]. However, as restricted using a substitute measure with selecting those regions and the scale, the individual's determinant of the hessian might make used for both.

Table 1: Equation for GLCM

1.	Energy	$F_1 = \sum t \sum f \{ p(t, f) \}^2$
2.	Contrast	$F_2 = \sum_{n=0}^{N-1} n^2 \{ \sum t \sum f \{ p(t, f) \} \}$
3.	Entropy	$F_3 = \sum t \sum f \{ p(t, f) \} \log(p(t, f))$
4.	Inverse Difference	$F_4 = \sum t \sum f \frac{p(t, f)}{1 + t - f }$

In 'interest points' T-junctions are chosen during dissimilar areas in the picture. Repeatability is the majority important property from claiming an investment purpose identifier furthermore which communicates the unwavering quality of a identifier under separate review states to finding those same physical interest focuses. Next, utilizing characteristic vector those neighborhood from claiming each premium perspective may be control. This descriptor need on make dissimilar also at the same time noise, geometric and photometric deformations like identification displacements. Finally, the middle of deferent pictures the descriptor vectors are matched dependent upon a separation at the middle of the vectors, e. g. euclidean separation. As stated by straight [19] SURF's descriptor identification is faster than the other existing methods. Bay, Ess What's more Tinne [6, 19] inferred that Hessian-based detectors need aid a greater amount of stability. They revised the harr based descriptors using the approximations to restrict the loss of precision.

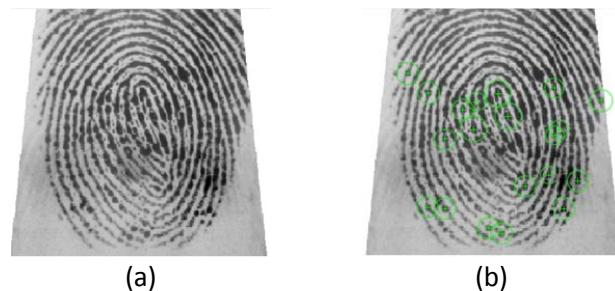


Figure 3: Detection of SURF feature points (a) Input Image (b) Interest Points

Gaussian might perfect to scale-space analysis, yet that should one gesture they must make destroyed which introduces artefacts, particularly over little Gaussian Kernels. Surf pushes those close estimation undoubtedly further, using the individuals box filters. These assessed second-order Gaussian

derivatives, Also Camus aggravate evaluated exceptionally using key analytics images, uninhibitedly for their measure. Box filters might procedure expansion close estimation of the Gaussian subsidiaries similarly there need aid vast number different wellsprings around tremendous noise in the get ready chain. Surf need been showed up to make more than five times speedier again distinction for Gaussian.

C. Gabor Wavelet

Wavelet transform provide a multi-resolution approach to texture analysis and classification [10]. Gabor wavelet proves to be very useful texture analysis and is widely used. Gabor Wavelets are group of wavelets in which each wavelet capturing the energy at a specific direction and frequency. So Gabor wavelet provides the local frequency description in images. Textures features can be extracted from these groups of energy distribution. The scale and orientation invariant property make Gabor wavelet to useful for constructing feature vectors [10] [15].

Gabor wavelet is the multi-scale and multi-orientation approach. The Gabor function is the Gaussian modulated by a complex sinusoid ω and the standard deviation σ_x and σ_y of the Gaussian envelop as follows [10].

$$\psi(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{[1 - (\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2})] + j2\pi\omega x} \quad (1)$$

Gabor Wavelet are obtained by dilation and rotation of generating function $\psi(x, y)$ as follows [15]:

$$\psi_{mn}(x, y) = a^{-m} \psi(x', y') \quad (2)$$

Where $x = a \cdot m(\cos\theta + y \sin\theta)$, $y = a \cdot m(-x \cos\theta + y \sin\theta)$, $\theta = n\pi / K$, $m = \{0, \dots, S-1\}$ and $n = \{0, \dots, k-1\}$, represents scale and orientation respectively . S and K are the number of desired scales and orientation respectively. The Gabor window with image I is given by [10],

$$G_{mn}(x, y) = \sum_s \sum_t I(x - s, y - t) \psi_{mn}(s, t) \quad (3)$$

Gabor wavelet accurately extracts the texture of the image but it required large computation because of large feature vector for each image.

Table 2: Differential Analysis

Technique	Advantage	Limitations
GLCM	<ul style="list-style-type: none"> • Small Dimension Feature Vector. • Covers the Direction Property • Less Computation Time • Multi-orientation & Robust 	<ul style="list-style-type: none"> • Only consider Gray scale images.
SURF [8]	<ul style="list-style-type: none"> • Invariant to Rotation and Scale detector and descriptor. • Images under different 	<ul style="list-style-type: none"> -Only consider Gray scale images. • High Dimension

	conditions such as low light images, blurred images are also recognized.	Feature Vector.
GBW [10][11]	<ul style="list-style-type: none"> • Extracted important components of images • Invariance to illumination, rotation, scale, transform 	<ul style="list-style-type: none"> • Loss of spectral information due to incomplete cover of spectrum plane

D. Similarity measurement

Euclidean separation may be the vast majority regularly utilized to similitude estimation on picture recovery due its effectiveness Furthermore viability. It measures those separation the middle of two vectors of pictures Toward figuring those square root of the aggregate of the squared supreme contrasts. It is calculated as [11]:

$$d = \sqrt{\sum_{i=1}^n (D_i - d_i)^2} \quad (4)$$

IV. RESULTS AND ANALYSIS

The experiments have been carried out on the standard datasets FVC2000, FVC2002 and FVC2004. The sample query image input and retrieved image is shown in figure 4. To evaluate the performance, precision and recall is calculated as follows:

Precision (P) = True Positives / (True Positives + False Positives)

Recall (R) = True Positives / (True Positives + Missed)

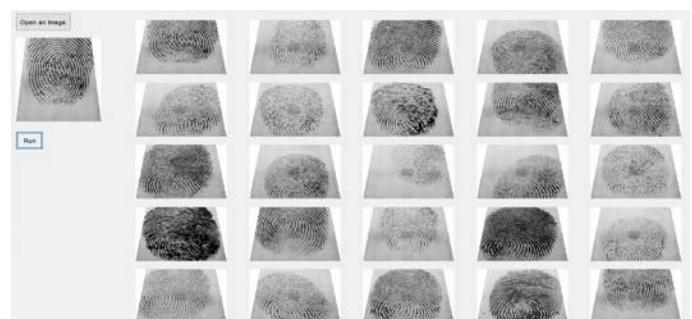


Figure 4: Retrieval Result

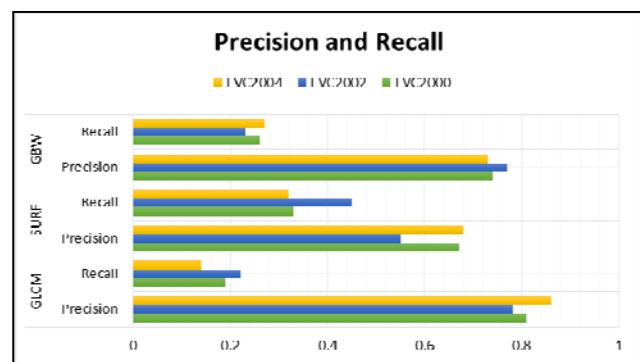


Figure 5: Precision and Recall

V. CONCLUSION

In this research paper different methods of texture based feature extraction for fingerprint image are discussed. From that it concludes that Gray-level co-occurrence matrix (GLCM) has fewer dimensions than Surf and Gabor Wavelet and also it is faster. Improved results are found using Gray-level co-occurrence matrix. Its gives 81.67% accuracy when Surf gives 63.33% and GBW gives 74.67%. The implantation results shows that texture features gives higher precision value compared to other features specifically the applications like pattern retrieval or recognition.

Table 3: Analysis

Datasets	GLCM		SURF		GBW	
	Prec- ision	Recall	Preci- sion	Recall	Prec- ision	Recall
FVC2000	0.19	0.81	0.33	0.67	0.26	0.74
FVC2002	0.22	0.78	0.45	0.55	0.23	0.77
FVC2004	0.14	0.86	0.32	0.68	0.27	0.73
Accuracy (%)	81.67		63.33		74.67	

REFERENCES

- [1] Dr. Ebtesam Najim Abdullah AlShemmary, “Classification of Fingerprint Images Using Neural Networks Technique”, Journal of Engineering (JOE), World Science Publisher, United States, Vol. 1, No. 3, 2012,
- [2] K. Jain, A. Ross and S. Prabhakar, “An introduction to biometric recognition”, IEEE Transactions on Circuits and Systems for Video Technology, 2004
- [3] Herbert Bay, Andress Ess, Tinne Tuytelaars, Luc Van Gool, “Speeded-Up robust features (SURF)” Elsevier:Computer Vision and Image Understanding, Volume 110, NO.3, 2008, pp. 346–359.
- [4] Csurka G., Dance C., Fan L., Willamowski J., and Bray C. “Visual categorization with bags of keypoints,” Workshop on Statistical learning in compute. Vis., ECCV. Vol. 1. No. pp.1-22. 2004.
- [5] Lowe D.G., “Distinctive image features from scale-invariant keypoints, “ Int. journal of compute. Vis., vol. 60, pp. 91–110. 2004.
- [6] Javier A. Montoya Zegarra, Neucimar J. Leite, Ricardo da Silva Torres, “ Wavelet-based fingerprint Image Retrieval”, Journal of Computational and Applied Mathematics, Elsevier, 2007
- [7] Bay H., Tuytelaars T., and Van Gool L. “Surf: Speeded up robust features“ Compute. Vision–ECCV, Springer, pp. 404–417. 2006.
- [8] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F,& Van Gool, L. “A comparison of affine region detectors“. International journal of computer vision, 65(1-2), pp. 43-72.,2005.
- [9] Mukherjee, J., Mukhopadhyay, J., & Mitra, P. “A survey on image retrieval performance of different bag of visual words indexing techniques” . In Students' Technology Symposium (TechSym), pp. 99-104. IEEE,2014.
- [10] J. Yue, Z. Li, L. Liu, Z. Fu, “Content-based image retrieval using color and texture fused features”, Mathematical and Computer Modelling 54 (2011) 1121–1127.
- [11] Hartigan, J. A., & Wong, M. A. “Algorithm AS 136: A k-means clustering algorithm”. Applied Statistics,pp. 100-108. 1979.
- [12] K. Velmurugan, Lt. Dr. S. Santhosh Baboo, “Content Based Image Retrieval using SURF and Colour Moments”, Global Journal of Computer Science and Technology, 2011
- [13] K Lu, Jidong Zhao, Yue Wu, “Hessian Optimal Design for Image Retrieval”, Journal of Pattern Recognition, ACM, 2011
- [14] Guerrero, M. “A comparative study of three image matching algorithms: SIFT, SURF, and Fast.“ Master Thesis (Utah State University).2011.
- [15] X. Ou, W. Pan, X. Zhang, P. Xiao, “Skin image retrieval using Gabor wavelet texture feature”, International Journal of Cosmetic Science, Vol. 38, issue 6, Dec 2016, pg. 607-614.
- [16] Pönnitz, T., Stöttinger, J., Donner, R., & Hanbury, “A. Efficient and Distinct Large Scale Bags of Words“. Austrian Research Promotion Agency (FFG). 2010.
- [17] Halawani A., Teynor A., Setia L., Brunner G., and Burkhardt H. “Fundamentals and Applications of Image Retrieval: An Overview, “ Datenbank-Spektrum, vol. 18, pp. 14–23. 2006.
- [18] Chang, R. I., Lin, S. Y., Ho, J. M., Fann, C. W., & Wang, Y. C. (2012). “A novel content based image retrieval system using k-means/knn with feature extraction“. Computer Science and Information Systems, 9(4), pp.1645-1661. 2012.
- [19] Tuytelaars, T., & Mikolajczyk, K. “Local invariant feature detectors: a survey“. Foundations and Trends® in Computer Graphics and Vision, 3(3), pp.177-280. 2008.
- [20] Miroslav Benco, Robert Hudec, “Novel Method for Color Textures Features Extraction Based on GLCM” Radio engineering, Vol. 16, NO. 4, December 2007.
- [21] Khater Meshkini and Hassan Ghassemian “Texture classification using Shearlet transform and GLCM”, International Conference Electrical Engineering (ICEE), 2017 (IEEE Xplore)
- [22] Sheshang D. Degadwala & Dr. Sanjay Gaur “Privacy Preserving System Using Pseudo Zernike Moment with SURF & Affine Transformation on RST Attacks” Vol. 15 No. 4 April 2017 International Journal of Computer Science and Information Security.
- [23] Sheshang D. Degadwala & Sanjay Gaur “A study of privacy preserving system based on progressive VCS and RST attacks” International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC) 2016.

PROACTIVE DETECTION OF CATASTROPHE TRENDS FOR RESCHEDULING REAL-TIME SYSTEMS WITH SCENARIO SHIFT

A. Christy Persya¹, T.R.Gopalakrishnan Nair²

¹Advanced Real-Time Computing Group, RRGI Research Centre, Bangalore, India.

²Advanced Real-Time Computing Group, RRGI Research Centre, Rector RRGI & Visiting Professor,
NIAS, Bangalore, India.

christypersya@gmail.com,trgnair@gmail.com

ABSTRACT

The objective of this paper is to establish the fact that a complex real-time system can be observed for a possible catastrophe shift at an early stage and impose viable rescheduling strategies that can help in realizing the best management of scenario, including the stability of the system under threat. Identifying and eliminating the threats leading to scenario shift are an essential part of the engineering of complex systems. The real-time systems are complex systems assuring the temporal correctness of the inputs and the produced output. The regular test on sensors and other components of the system searches for the predefined patterns of failures. We propose to improve and augment such detection capabilities of the monitoring systems for better reliability. A catastrophe shift detection mechanism that can offer an early warning of the possibility of a catastrophe and it can eliminate the occurrence of an unexpected event. The catastrophe bifurcation model applied over the system parameters can provide sufficient indication of the sequence of events that may lead to a failure of the system. The approach is based on general catastrophe theory that allows detection of unexpected changes in systems, and the estimation of its associated probable severity.

KEYWORDS: Catastrophe theory, Catastrophe monitoring, Catastrophe trend detection, Real-time systems, Monitors.

1. INTRODUCTION

During the twentieth century, computer controlled systems were widely accepted in all domains of industrial activity ranging from nuclear power plants to avionics systems of common airplanes as well as space shuttle systems. These complex systems are mostly managed by embedded systems with dedicated functions within a framework of larger mechanical or electrical systems. They often remained with real-time computing constraints. These real-time computing systems not only depended upon the correctness of output but also depended on the time mark of the delivery of the output. Certain unexpected and expected behavior patterns used to bring failure in the system, and these led to several modes of catastrophic situations. The uncertainties related to these failure modes can

occur in any one of the three domains such as time, space, and its unique combinations. The uncertainty in time can be well understood with the knowledge of unpredictability of response times. The hard real-time system is expected to schedule and execute all assigned tasks before its deadline to avoid catastrophe. In certain times, it has to take care of the most critical tasks to execute by suspending few low priority tasks [21]. In hard real-time domains, such as avionics, air traffic control, medical monitoring, and power plant control systems, they must generate control actions (tasks) to meet deadlines, or catastrophic failure might occur. Control systems operating in real-time environments must not only choose appropriate actions in varied situations but also must act at appropriate times. Conventional real-time system framework yields fairly rigid systems that cannot easily move on to a sudden shift in changing scenario and associated requirements. Hard real-time systems depend on the programs and the schedulability analysis of prediction algorithms. Most modern real-time applications like space platform management, high speed autonomous system manures or fully automated critical safety systems require a fair set of new capabilities to handle the demands of the increasing scale, complexity, and other external factors. These systems need to be much more flexible and intelligent in dynamic and complex situations to provide guaranteed behavior and avoid catastrophe. Real-time systems (RTS) have only finite time to sense its environment and compute its dynamic response. They are, in general, designed to shift from one operational mode to another based on predefined operations or stages [1, 2]. This underlines the need to develop intelligent real-time systems that will allow predicting failure detection at an early stage in order to maintain the acceptable performance of the system with some probable unexpected events.

The last few decades witnessed the development of many technology areas that employ real-time systems such as aerospace vehicles ranging from ordinary aircraft to long-term interplanetary transit modules as well as several chemical and nuclear process installations threatened by unperceived scenarios. All of these applications demand more “intelligent” schedulers for an “intelligent” real-time system design. This, in turn, is in need of suitable models of predicting run away bifurcation phenomena as in catastrophic systems.

The paper is structured as follows. Section 2 discusses existing monitoring mechanisms that are currently prevalent in RTS and elaborates on the need for scenario shift detection. Section 3 addresses the cusp catastrophe monitoring and detection system in a sample RTS. Section 4 concludes the paper.

2. RELATED WORK AND MOTIVATION FOR SCENARIO SHIFT

2.1 Motivation

Crucial components in a real-time system comprise of sensors, control units, schedulers, processors, actuators, etc. These components in the system undergo various testing phases and online-monitoring mechanisms for fault detection. Many times, if the most critical hard deadline task misses its deadline during execution, it might lead to being a cause of a catastrophe. In the United States, a partial nuclear meltdown occurred in one of the two Three Mile Island (TMI) nuclear reactors in 1979 [3]. The accident began with some failures in the non-nuclear secondary system which then triggered another event in the primary system. Finally, it allowed large amounts of the nuclear reactor coolant to escape and caused the catastrophe. After the TMI accident, one of the causes identified for the

failure as described in a survey report [3] was related to the wrong expectation exercised by the operators related to a command given for valve closure.

Chernobyl disaster in 1986 is another catastrophic accident that happened in a nuclear power plant [4]. This accident occurred because of a sudden and unexpected power surge. A series of steam explosions and a reactor vessel rupture occurred when an exponentially large spike in power output was caused because of attempting an emergency shutdown. Fukushima Daiichi nuclear disaster in 2011 resulted in a nuclear meltdown of three of the plant's six nuclear reactors [5].

Similar kind of catastrophe would have occurred in aircrafts such as the SilkAir Flight 185 in 1997 and Air France Flight 447 (AF447/AFR447) [6] [7].

The Australian Transport Safety Bureau conducted an analysis of a problem raised from an error in the Air Data Inertial Reference Unit (ADIRU) software which was exposed by a series of events and reported that the bug was not revealed during certification testing [8]. Hence there are definite probabilities existing in working systems to have a drift in performance which can lead to a catastrophe.

In all the above cases, the catastrophe was caused by a series of distinct events with unusual features; wherein the system would have been in a regular mode even a few minutes before the catastrophe. Before the safety handlers could have been enabled, the system broke down, and catastrophe took place. Thus, the current practice of relying on periodic surveillance tests alone to check the health of hardware, and software is found to be clearly insufficient. Hence we need to move forward in systems design approaches incorporating various modes of intelligence which will monitor the systems much beyond period surveillance with set point control.

Besides the initial schedule, the intelligent RTS must be capable to learn the variations that are happening in the parameters that monitor the environment change, perceive the changes, the reason for the failure causing events and propose appropriate actions in a completely automated mode. A single point failure does not cause a catastrophe rather it is built up by multiple failures, mostly in a minute fraction of the time. There are many arguments that state that degraded conditions must be predicted as much earlier as possible to avoid catastrophe. The design of the critical, complex real-time systems needs integration of certain intelligence to deal with such conditions. Such intelligence is required to reschedule the tasks automatically by allowing the unexpected arrival of critical tasks for execution. One of the principles of modern approach is that the theory needs to be revised when observations do not match the calculated predictions [9]. The identified cause for the catastrophic failures and motivation for the proposed model is the failure of emergency planning and systemic failures. The cause of all the failures looks similar in means of underestimated risks. If catastrophe cause is detected much prior by observing the catastrophe-prone variables and rescheduling is planned, the severity of the catastrophe can be reduced. Few existing online monitoring mechanisms and fault detection schemes are discussed below.

2.2 Related work

The existing detection mechanisms can be grouped into four categories. They are sequential offline check, online monitoring, artificial intelligence monitoring, and model-based monitoring along with fault detection. Regular checking pattern followed in traditional RTS has a sequential check on memory, processors, engine and all components during system startup. An example of the sequence is as follows: i) Processor will be checked ii) Program will be checked for accuracy iii) RAM will be checked iv) Watchdog timer will be verified. In this entire process, there is no preemption, so no task context switch takes place. The worst case execution time (WCET) would be more than the input data rate. So sequential tasks in a predefined manner get executed before the deadline in simplest modes of operation. Fault tolerance also may be brought in by enabling duplex, quadruplex or higher order multiplex systems. In such redundant configurations, any channel failure may enable the assigned redundant channel to take over and salvage the failure mode.

The challenges in online monitoring include implementing efficient monitors that are synthesized from high-level behavior specifications.

In [10], a new monitoring system using object oriented concepts and artificial intelligence (AI) was introduced. During a disaster occurrence, operators cannot take the most correct decision based only on the status output of alarms, indicators. As a part of the progress in this area, the computer systems were introduced to monitor the real-time performance and help the operator in decision making. Run time monitoring, and verification was incorporated to predict the changes [11] [12]. Certain AI based monitoring system has been effective since 1997. In an artificially intelligent monitoring system (AIMS) [12], the acquired and calculated variables with their interdependencies are mapped into a hierarchical objects network. The state of monitored variables updates a fact-base which is used to activate the knowledge base rules.

The model-based monitoring system uses a nonlinear state estimation technique coupled with a probabilistic based statistical hypothesis test. The fault at early times are detected and identified from the sensors and, other components of the RTS and changes in the stochastic characteristics of measured signals [13].

3. DETECTION OF CATASTROPHE TRENDS

3.1 The catastrophe theory

Catastrophe theory is a branch of mathematics for dynamical systems [14] that was proposed by a French mathematician Rene Thom in 1960s. The theory is characterized by sudden shifts in behavior arising from small changes in circumstances. This approach has been applied to attitudes and stage transition proposed by Piaget and various other dynamic systems [15].

The singularity formulation was first presented by Thom in 1972 and it was widely discussed and formalized further with examples. Poston and Steward in 1978 have made the subject very clear with the clear picture of the dynamical system and the elementary catastrophes [15]. In 1976 Zeeman renamed the theory as catastrophe theory and

explored its applicability to a variety of applications [16]. In this, Gilmore explained the Moores' stability law and the various catastrophe flags for sudden jump and discontinuities were discussed here.

The goal of the catastrophe theory is to specify a set of common criteria for discontinuity phenomenon. It classifies the various ways in which a system can undergo sudden large changes in behavior as one or more of the variables that control it are changed continuously and simultaneously [17]. In the recent past, there were several attempts to apply this theory to predict the unexpected behavior in biological systems and other artificial systems [17]. Though the formulation of a mathematical model for catastrophe is complex, it can give handsome rewards in physical systems too as it is explored hitherto.

3.2 Proactive strategy

The conventional RTS are commonly designed to accommodate fairly well known tasks with finely estimated execution times within the scope of processor speed, memory access and associated interactive delays of various components. In hard core RTS, the assurance built into the scheduler is proven sufficient enough to manage the roll out of tasks of each frame producing mostly deterministic results enabling the plant to run in a predetermined way. There are several factors affecting a plant which puts the deterministic RTS into arbitrary positioning with respect to the safety and security of a plant which is not very unusual. This is true in the case of critical RTS employed usually in aerospace systems, nuclear operations, various high sensitivity chemical plants and warfare systems. However, the response to such rare but possible scenarios from fully pre determined approach of RTS is increasingly found insufficient. It calls for a proactive strategy and algorithmic approach to foresee catastrophic trends from the variable spectrum available for the hard real-time purposes. This proactive positioning of the real-time monitoring and management of complex systems can be achieved through catastrophe models and it's theoretical approaches and it is described here.

3.3 Cusp Catastrophe Monitoring And Detection System [C²MDS]

As it is evident from the previous discussions, the system may have to reconfigure to incorporate catastrophe theory to take precautionary measures if catastrophe prone variations are observed in the plant under control.

The next generation RTS design mostly will contain a fully proactive set of components looking into the possibility of disasters or catastrophe. Each component can be a function in an RT module and will have dependency on N variables. Among the N variables of the system it is possible to extract a set of control variables, and behavioral variables required for catastrophe assessment. It can be described by the system designer based on the input requirements of the catastrophe model developed for the plant. While the control variables are independent variables, the behavioral variables are dependent variables whose variation in the behavioral plane is dependednt upon the two control variables as shown in Fig 2. In catastrophe models [17], the catastrophe effect like sudden

jump happens in the behavior plane when the control variables are in bifurcation set. Consider a scenario related to aircraft flight where there is a need for a maneuver to compensate and avoid possibilities of high structural shock from adverse spatial clouds. There is a definite probability that due to high wind (air turbulence) the control system may call for extreme maneuver of the body resulting in structural challenges. In the case of linear model, as shown in Fig.1, the internal control error accumulation can drift the control system to an instability mode [18]. Similarly, the environmental challenge can impose a system to drift to high risk mode. In a situation where the system is moved to instability mode and a challenge happens from the environment, it can be seen that the instability mode cannot flip over to high risk mode traversing the neutral point. Hence, a better theoretical approach is required to produce proactive plant protection scheme.

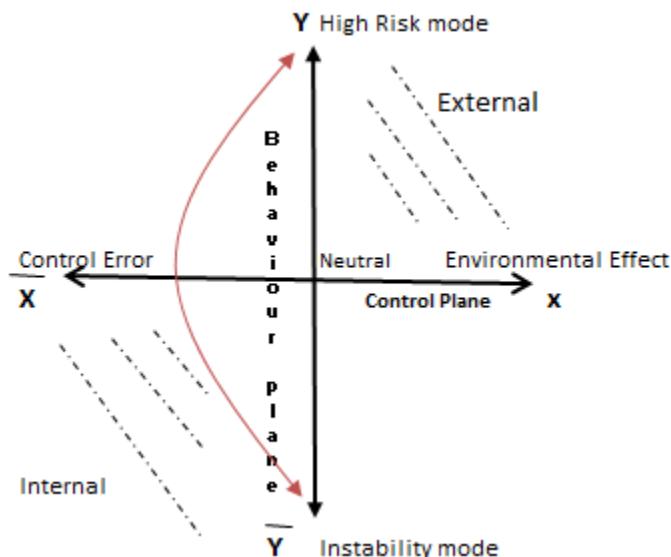


Fig. 1. Linear model in RTS

Linear model explains the sudden changes from instability mode to high risk mode as a function of the characteristics of environment and control error. The behavior variable ranges from instability mode to high risk mode which is controlled by the control error and/or environment. The variation in the control plane depicted in the $X\bar{X}$ axis will be expressed along $Y\bar{Y}$ through high risk mode and instability mode [18].

If the system state over the $X\bar{X}$ axis tends to be closer to the neutral point, there exists neither high risk nor any instability. From this point, an increase in environmental effects or control error leads to a continuous increase towards instability or high risk as far as aircraft is concerned. However, if air turbulence level is increased for an aircraft in which there exists a high control error already, it can lead to catastrophe as per the catastrophe theory. Same type of catastrophe tendency exists in the reverse way. The magnitude of catastrophe positioning depends on the distance from the neutral point in $X\bar{X}$ axis for both environmental and control error challenges.

There can be simultaneous challenges of control error being at maximum point and environmental challenge drifting to its own maximum. In such situation, there will be unexpected transitions from instability mode to high

risk mode leading to oscillations between them. In this situation, the linear model is insufficient to depict the transition along YY' because the actual path of transition cannot be shown in the linear map and it may traverse curved paths which cannot be easily analyzed. Hence improved models are required to deal with such behavioral occurrences. The catastrophe model involving cusp geometry is one of the best alternatives [18].

The two transitions which lead to catastrophe as represented in the linear model posses a discontinuity in the path that cannot be shown in a linear model. This discontinuity that cannot be shown in a linear model used in conventional real-time systems can be effectively represented in a cusp catastrophe model which supports discontinuities.

A cusp model which can depict sudden discontinuities is shown in Fig. 2.

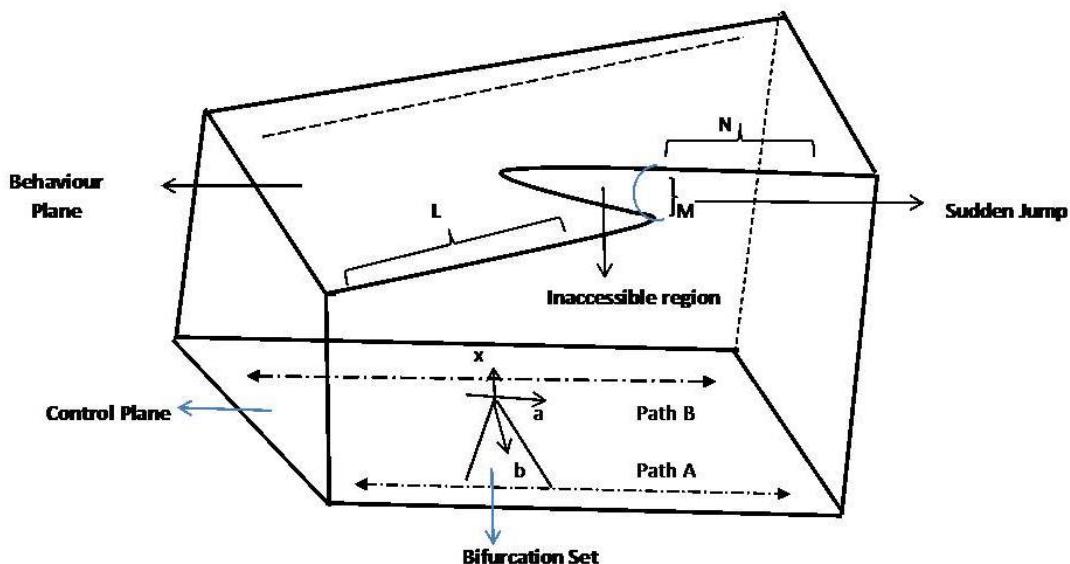


Fig. 2. The cusp model in RTS.

In Fig.2, a and b are the control variables (independent) of catastrophe formation, like control error and environment effect of the system. The behavior plane (dependent) has instability mode, and high risk mode. The dotted line in the behavior plane is in correspondence with the path B shown in the bottom control plane. Transitions can take place in the path A, and path B. Path B in the behavior plane does not have the possibility for bifurcation because the path on the control plane is struck to one situation only like maximum control error or environmental effect. Path A depicts another example where there is a possibility for the control plane to be affected simultaneously by the control error and environmental effects. The figure shows the three segments path A has traced because of the simultaneous control plane demand occurred from the system control error and the environmental challenge. The segment N shows the normal behavior where the system will be having proper characteristics for systems control. The region M shows where two control regions simultaneously exist farthest

from the neutral point. This region shows the possibility for the sudden jump from normal behavior to deviated behavior. The control path has got several bifurcation points.

If we allow the bifurcation to widen, it leads to behavioral path tracing much widened behavioral change usually called inaccessible regions. These inaccessible regions indicate the plant has undergone deviation leading to behavioral change which cannot be retraced and the plant still holds the challenge situation existed in the control domain. The control domain has got two challenges. Now the plant has shifted its focus from the previous challenge to the second one. It is evident from this scenario that the previous challenge still exists and it cannot go back such a way that it will be able to control that one. This is a serious situation, and it calls for emergency intervention from the external world indicating a tendency to catastrophe has started.

Region M contains those inaccessible regions which could create trouble for the plant control. In the case of a real-time system, this is the most vital point where an identification of emergence of bifurcation can be detected. The detection does not mean that already catastrophe has happened. But it positively indicates the definite probability of the plant slipping into a situation of less controllability.

The catastrophe brings in a behavior of the plant in an uncertain region in which the system may not be able to identify the specific property of the plant. In this situation correct decision making becomes vague, and it calls for scenario shift for better management. The curve in the behavior plane near the neutral region is the inaccessible region which is the unstable state. So if control error and environment effect are simultaneously increased from the neutral starting point, the upper or the lower plane becomes the option as the central region is inaccessible region. The system will either be in instability mode or high-risk mode but will be unstable between these two. That is depicted in the region M. The M region indicates the discontinuity in the behavior. The region L shows the behavior of the system is changed abruptly from M to L indicating uncontrollable scenario. At this stage, the system will have the presence of catastrophe indicators like bimodality, hysteresis, and divergence as a result of discontinuity which is discussed later as in Fig 3.

3.3.1 Catastrophe Indicators

The presence of catastrophe is associated with the catastrophe flags. They are easy to recognize and provide information about the underlying catastrophe. There are eight catastrophe flags prescribed by Gilmore [18]. They are hysteresis, divergence, sudden jumps, inaccessibility, modality, bimodality, divergence of linear response, critical slowing down/mode softening, and anomalous variance. Each flag is a behavioral property that has been mathematically derived from the catastrophe theory [18]. The presence of one catastrophe flag in the system is an indication of other properties either already present or imminent. These eight flags were used in a practical environment of studying animal behavior in cognitive developmental research [19]. The first five occurs when there is a qualitative change in the system. The remaining three occurs when there is a qualitative change but also may be observed before a phase change.

Modality – In general, the system has distinct types of behavior. The upper surface and lower surface of the cusp catastrophe represents two different types of stable behavior. The intermediate part represents the unstable mode of behavior, which is modality.

Sudden Jump- The system may suddenly jump from one mode to another mode in the behavior plane as the control variables vary. These jumps represent the transition from one local minimum to a global or another local minimum.

Hysteresis- A sudden jump from one mode to another which leads to catastrophe occurs in the bifurcation plane. The reverse process might also occur in the same control variable.

Divergence- When the system moves from one mode to another, the final state of the system is in evolution. The final state depends on the upper sheet of the cusp model. It means that it depends on the initial conditions of the system.

Inaccessibility- There is a saddle point (point of inflection) in between two stable points. This region during the mode change is inaccessible which is shown as a saddle point.

Bimodality divergence of linear response- This means variables in the neighbourhood of the bifurcation area show large fluctuations after reaching the stable or settling point. After some oscillations, it might stay in the initial state or a new state.

Critical slowing down/mode is softening- It means that, after a perturbation, it takes a certain amount of time before stable behavior returns.

Anomalous variance- This is an increase in the variance of behavior that occurs in the neighbourhood of the bifurcation set (e.g., in the area where sudden transitions are possible). In this area, we can expect large fluctuations in behavior.

The potential function of cusp with two control variables and one behavior variable is given as

$$V_{ab}(x) = \frac{1}{4}x^4 + \frac{1}{3}ax^3 + bx \quad (3.1)$$

The balanced surface of $V_{ab}(x)$ is

$$x^3 + ax + b = 0 \quad (3.2)$$

The singular point of $V_{ab}(x)$ is

$$3x^2 + a = 0 \quad (3.3)$$

Critical points occur when the graph of a function f has a horizontal tangent. The critical point is any value whose

function is not differentiable, or its derivative is zero. It can be derived by putting the derivative $\frac{dy}{dx}$ equal to zero. For the given points in control plane along the path, a and b is given in Fig. 3.

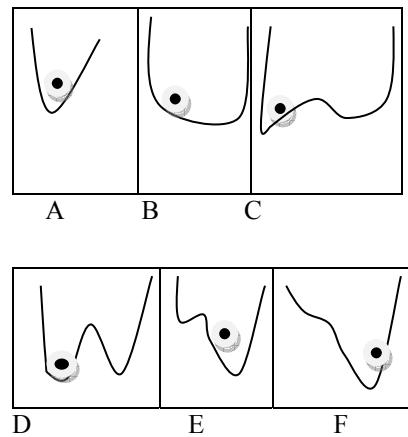


Fig. 3. Continuous increase leads to sudden shift

For a given point in the control plane, A shows the local minima and B to E shows the smooth transition that had a local maximum in C and inaccessibility region as a raised point between two modes in D. E shows the sudden jump. The transition from A to F shows the hysteresis in the form of sudden jumps when leaving the bifurcation set. The famous figure of cusp catastrophe is shown in Fig 4.

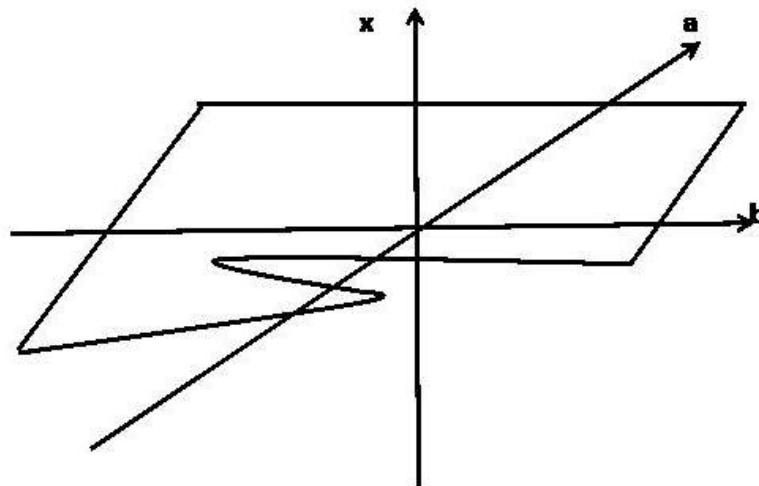


Fig. 4. Cusp model

The two different behaviors of the quartic function are separated by the curve shown in Fig 4.

$$8a^3 + 27b^2 = 0 \quad (3.4)$$

When $8a^3 + 27b^2 < 0$, two minima exist; in the case where inequality is reversed, there is a single minimum. The point on the curve of Fig 5 is called a cusp, and hence this catastrophe is called a cusp catastrophe.

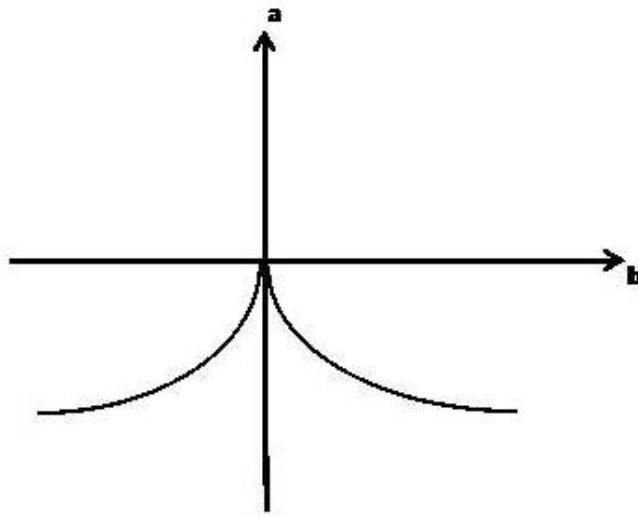


Fig. 5. Cusp Point

3.3.2 Modelling A Simple RTS With Catastrophe/Bifurcation Detection

The construction of cusp model starts with the identification of variables that control the transition. This model shows discontinuity in behavior variables as a function of continuous variation in control variables.

Here, apart from catastrophe modelling, the first step is catastrophe detection which is based on Gilmore work [19]. This involves the identification of typical properties of behavior that indicates the presence of catastrophe. The second step is the catastrophe analysis which consists of a mathematical analysis of the dynamic equations of the transition process. So the catastrophe analysis requires knowledge of mathematical equations related to the transition process. The main challenge here is the actual fitting of the cusp model to the data provided by the system designer. Considering two assumptions,

- 1) A local failure does not cause the immediate collapse of the entire structure.
- 2) The entire structure fails immediately after one of its critical components fails.

In the hard RTS which is under consideration, we have the scheduler with preemptive scheduling classifying task based on its priority and time for completion with EDF. This scheduler in effect will be classifying the task and dispatching based on hard RT rules. Several channels of tasks will be formed connecting to subsystems which need to execute this. Hence, subsystems S_1, S_2, \dots, S_n also gets segregated according to the schedule. For every device under command will have a priority queue. So the corresponding devices or subsystem of a queue will be emptying this under the predetermined design accuracies. For a perfect working condition design, the queue will be at a negligible level of 0 or 1. If the task is categorized with soft RT condition, then a queue may build up and subsystem work as servers of the queue [20]. Any under performance of the subsystem will be reflected in the

service of the queue. There can be n queues uniquely coupled to a subsystem S_1, S_2, \dots, S_n or one queue can cater more than one subsystem.

Whenever the underperformance of S_1, S_2, \dots, S_n happens, the queue will be building up. In the case of hard critical channels, build up above one or two can lead to system troubles. So more than two becomes an invitation to trace paths of the point of inflexion.

In order to model, the catastrophe system, we have to consider the three conditions.

- 1) Perfectly normal working (x)
All subsystems and all queues are fully served.
- 2) Slightly underperforming situation (y)
A group of subsystems building queues within a tolerable limit in the critical and soft critical domain.
- 3) Development of tendency to fail (z)

Some of the queues will be building to the extreme level of server, and in most cases, this will be served by subsystems properly. Hence, this situation is sensitive to failure and as per the catastrophe theory the path through the neutral point which may be ready to bifurcate. In this stage, the queue continues to build without the tendency to reduce. That means subsystems are not capable of offloading. This is the failure mode.

Next is the derivation of the catastrophic performance with respect to subsystem performance when Earliest Deadline First (EDF) is scheduled. We use queuing theory based results in analyzing the scenarios of catastrophic performance. EDF is an optimal scheduling algorithm on preemptive uniprocessors. With scheduling periodic tasks that have deadlines equal to their periods, EDF has a utilized bound of 100 %. Thus the schedulability test for EDF is

$$U = \sum_{i=1}^n \frac{C_i}{T_i} \leq 1 \quad (3.5)$$

where the $\{C_i\}$ are the worst-case computation times of the n tasks and the $\{T_i\}$ are their respective inter-arrival periods. EDF will guarantee that all the task deadlines are met provided that the total CPU utilization is not more than 100%. Compared to fixed priority scheduling, EDF can guarantee all the deadlines in the system at higher loading. The performance of a real-time queuing system is measured by its ability to meet the deadlines of the tasks. This is in contrast to ordinary queuing systems in which the measure of performance is often task delay, queue length or utilization of a service facility.

We assume that the subsystems have exponentially distributed inter arrival and service times (M/M/1 queue). These hypotheses, which are very common to allow for the solvability of analytical models, cannot always be a faithful representation of the real world. We are demonstrating how effectively this can be done in a sample RTS in this

paper. In this model, the queue length can have a maximum allowed no of tasks such that the EDF schedulability test satisfies.

Let's assume

- i) The maximum allowed no of tasks as \bar{Q} .
- ii) The queue length as Q_{max} .
- iii) The mean arrival rate is $\frac{1}{\lambda}$.
- iv) The mean service rate is μ .
- v) The performance measure (P) of a uniprocessor queue is $P = \frac{\lambda}{\mu}$.
- vi) The time interval between checks is T.
- vii) The number of subsystems is $N = x + y + z$ where x is the normal working system and y is the slightly underperforming and z is the systems with the tendency to fail.

The drift equation is based on the probability that ($Q_{\text{max}} \geq \bar{Q}$). The queuing theory formulas used are only valid if the transient phenomena due to a change in the mean arrival rate of tasks are exhausted; this requires for a large value of T. In order to get the drift and potential function, we have to obtain the expressions of the mean recovery rate $\alpha(x)$ and the mean collapse rate $\beta(x)$ for the system.

The recovery rate $\alpha(x)$ at time t corresponds to the average recovery process taking place instantaneously in between the checking time T. According to the hypothesis α is given by the following for above said 3 cases:

Case 1: No queues are pending.

Case 2: Under secondary performance area, there can be a set of few numbers which can work in the underperforming area and some in the failed mode.

Case 3: The number of failed modes due to overstress may build up the queues.

The state equation can be given as $N = x + y + z$.

If $z = 0$, then $N = x + y$. So the system will be working.

If $z \neq 0$, then there is some critical problem.

$$\alpha(x) = \frac{N - x}{T} \quad (3.6)$$

where $N - x$ is the number of overstressed subsystem in the RTS at time t.

The collapse rate $\beta(x)$ at time t corresponds to the average number of working subsystems collapse due to over stress, and the same working subsystem fails more than once.

$q_{us} \geq \bar{q}$ indicates that for few cycles, few underperformance happening with the subsystem.

Therefore, P [underperformance happening] depends on $P[q_{us} \geq \bar{q}]$.

The length of the queue will be the sum of q_{sc} and q_{pc} , where q_{sc} is the scheduler count and q_{pc} is the pending count.

i.e., $q_{us} = q_{sc} + q_{pc}$

$$P[q_{sc} + q_{pc} \geq \bar{q}] = \rho^{(q_{sc}+q_{pc})} \quad \text{if } \rho < 1 \quad (3.7)$$

Because in M/M/1 model,

$$P_n = (1 - \rho) \rho^n, \quad n=0,1,2 \quad (3.8)$$

$$P[q_{sc} + q_{pc} \geq \bar{q}] = 1 \quad \text{if } \rho \geq 1 \quad (3.9)$$

For the analysis purpose, it is assumed that the subsystem failure nature is confined to a band of time T in which it may go in an underperforming way and necessarily return back within the time T. ρ will become 1 when the scheduler output per unit time or RT frame is just matching the total no of tasks executed by the total no of subsystems.

i.e., $\lambda = \mu.x$

$$\frac{\lambda}{\mu.x} = 1$$

Hence $\frac{\lambda}{\mu.x} = 1$

At any instant, when λ is becoming greater than $\mu.x$, it will lead to an occurrence of catastrophe.

When $\lambda > \mu.x$ system expansion is done.

$$\beta(x) = \frac{x}{T} \cdot \left[\frac{\lambda}{\mu.x} \right]^{q_{sc}} \cdot \left[\frac{\lambda}{\mu.x} \right]^{q_{pc}} \quad \text{if } \rho < 1 \quad (3.10)$$

$$\beta(x) = \frac{x}{T} \cdot 1 \quad \text{if } \rho \geq 1 \quad (3.11)$$

We assume that $x(t)$ is the total no of working subsystems out of n which are ready to execute and remaining are under overload and is capable of making a coming back in the next frame.

So,

$$drift(x) = \alpha(x) - \beta(x) \quad (3.12)$$

Substituting $\alpha(x)$ and $\beta(x)$ in the drift function,

$$= \frac{N-x}{T} - \frac{x}{T} \left(\frac{\lambda}{\mu \cdot x} \right)^{q_{se}} \cdot \left(\frac{\lambda}{\mu \cdot x} \right)^{q_{pe}} \quad \text{if } \rho < 1 \quad (3.13)$$

$$= \frac{N-x}{T} - \frac{x}{T} = \frac{N-2x}{T} \quad \text{if } \rho \geq 1 \quad (3.14)$$

The potential function of the drift equation while $\rho < 1$ and $\rho \geq 1$ is given below:

$$V(x) = \frac{1}{T} \left\{ Nx - \frac{1}{2}x^2 - \left(\frac{\lambda}{\mu} \right)^{q_{se}+q_{pe}} \left(\frac{x^{1-(q_{se}+q_{pe})}}{1-(q_{se}+q_{pe})} \right) \right\} + c \quad \text{if } \frac{\lambda}{\mu \cdot x} < 1 \quad (3.15)$$

$$V(x) = \frac{1}{T} \{ Nx - x^2 \} + c \quad \text{if } \frac{\lambda}{\mu \cdot x} \geq 1 \quad (3.16)$$

To have the continuity of $V(x)$, the potential function becomes

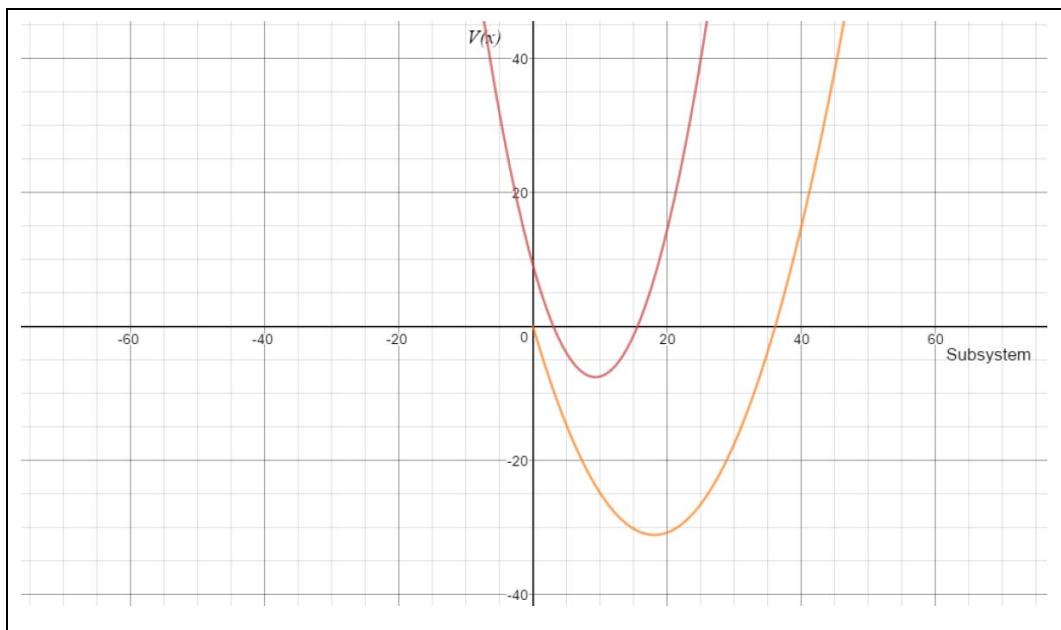
$$V(x) = \frac{1}{T} \left\{ 2Nx^2 - \frac{3}{2}x^4 - \left(\frac{\lambda}{\mu} \right)^{q_{se}+q_{pe}} \left(\frac{1}{1-(q_{se}+q_{pe})} \right) x^{1-(q_{se}+q_{pe})} \right\} + c \quad (3.17)$$

By analyzing the potential function with the simulation model, we have observed that it has a quartic shape similar to that of the cusp catastrophe model [22].

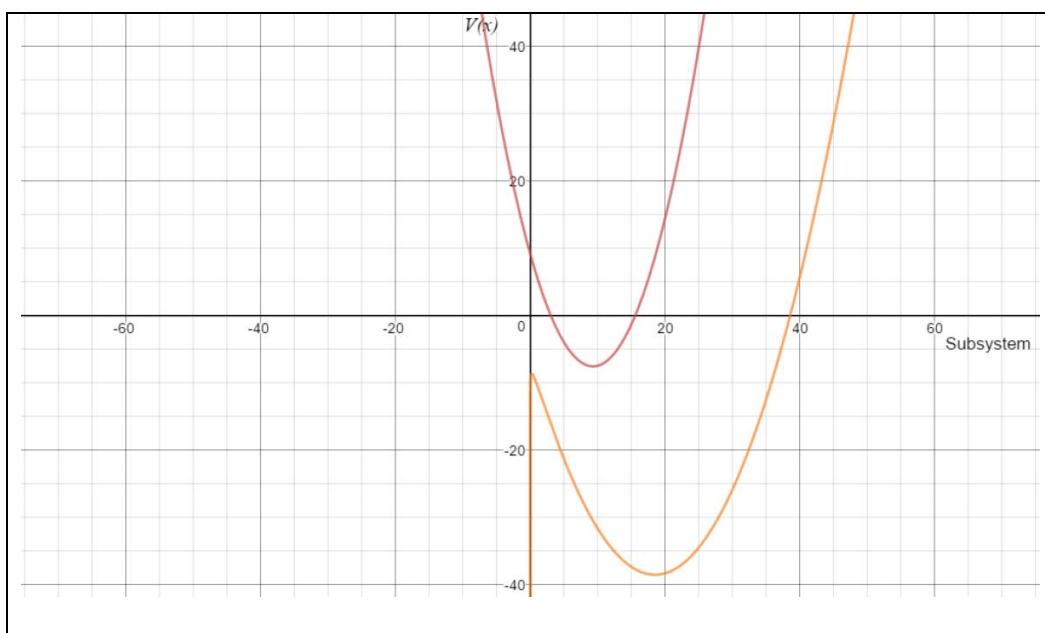
3.3.3 Study Of Emergence Of Catastrophe Indicators

The graph for the potential function is given Fig 6. The performance of potential function here is similar to the cusp model performance. The control parameters here are the ~~Ques~~ and the arrival rate with the behavior variable x as a number of working subsystems.

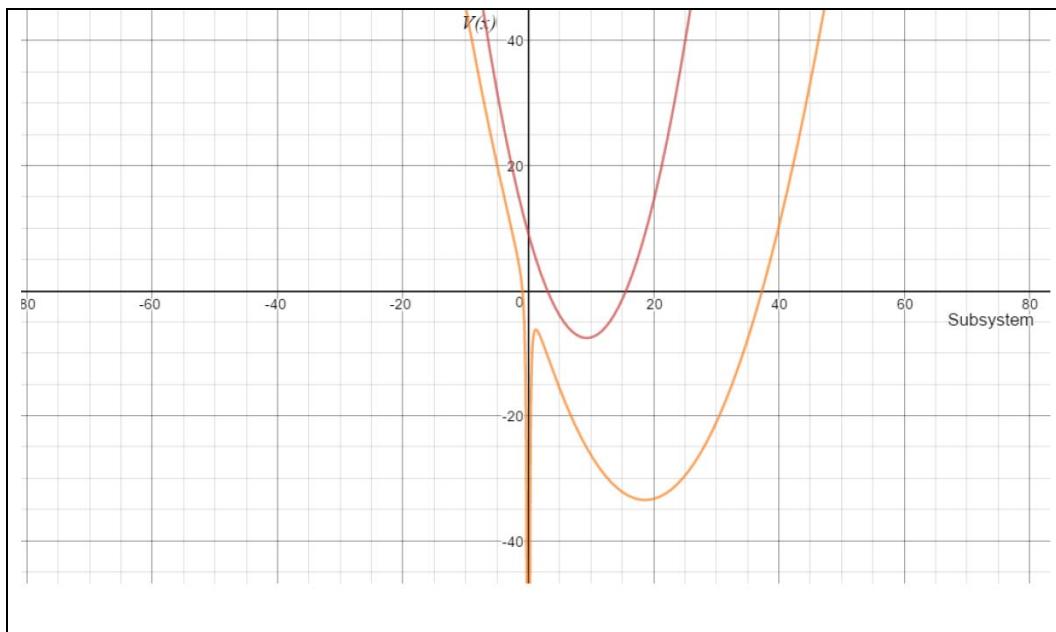
1. It has one minimum with one state.



2. It has one minimum with a point of inflexion.



3. It has two minimum and one maximum.



4. The new maximum disappears and leads to a new state.

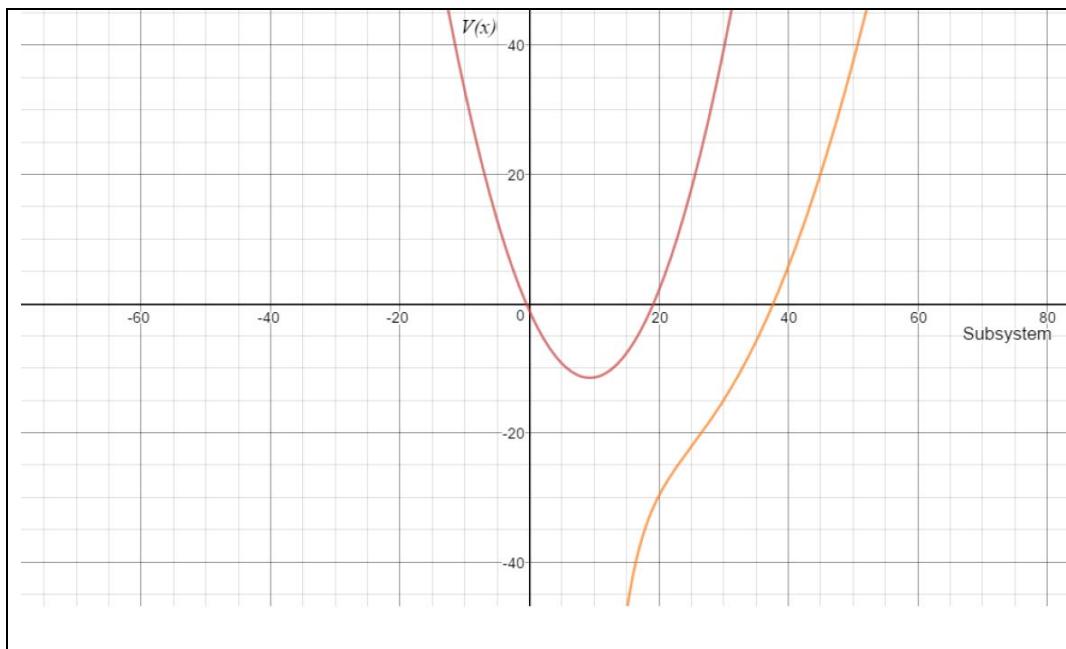


Fig. 6. Potential function performance

When the control parameter is with one value, the system performance is in one state. When the control parameter varies, this leads to a sudden jump to a new state. Thus, a sudden jump in behavior from one state to other state led by a small change in control parameter is defined as a catastrophe. Here, when the control parameters arrival rate and the queue length parameters vary, a drift happening in the stability behavior of the systems is observed.

4. CONCLUSION

Conventional real-time systems have designs applied to large scale systems and they form systems usually performing within a rigid architecture of timing constraints to keep up the system specifications. But there could still be a problem persisting such as the lack of capability of the real-time systems to detect trends and failure warnings based on authoritative models of system deviation. The field of modeling provides an effective theory of dealing with deviations in performance using catastrophe theory as a viable tool. The hybridization of this tool with existing contemporary models of real-time systems paves the way for introducing catastrophe analyzing schemes in critical plants. Here, we present an implementation of cusp model of catastrophe theory in a real-time system successfully. It also demonstrates the effective detection of the emergence of probably irrecoverable phases of the plant much before the real occurrence of the event which may be due in short time. It engages a quartic cusp potential function which is derived effectively from the properties of a typical real-time system. This real-time system contains a host of channels of subsystems which are engaged by an optimized scheduler issuing a sequence of tasks based on predetermined algorithms and priority distributions. This type of system design demands a detailed analysis of task characteristics of the real-time system under consideration. This process can lead to the development of the main thread of catastrophe design like the potential function. It can lead to a modern design of vigilant and autonomous real-time system.

REFERENCES

- [1] Real, J., & Crespo, A. (2004). Mode change protocols for real-time systems: A survey and a new proposal. *Real-time systems*, 26(2), 161-197.
- [2] Burns, A. (2014, December). System mode changes-general and criticality-based. In Proc. of 2nd Workshop on Mixed Criticality Systems (WMC) (pp. 3-8).
- [3] Rogovin, M. (1979). Three Mile Island: A report to the commissioners and to the public (No. NUREG/CR-1250 (Vol. 1)). Nuclear Regulatory Commission, Washington, DC (USA).
- [4] Shestopalov, V., Bohuslavsky, A., & Bublias, V. (2014). Groundwater vulnerability: chernobyl nuclear disaster (Vol. 69). John Wiley & Sons.
- [5] Holt, M., Campbell, R. J., & Nikitin, M. B. (2012). Fukushima nuclear disaster. Congressional Research Service.
- [6] ‘SilkAir Flight accident’, https://en.wikipedia.org/wiki/SilkAir_Flight_185
- [7] ‘Air France Flight accident report’, https://en.wikipedia.org/wiki/Air_France_Flight_447
- [8] Goodloe, A. E., & Pike, L. (2010). Monitoring distributed real-time systems: A survey and future directions.
- [9] Lala, J. H., & Harper, R. E. (1994). Architectural principles for safety-critical real-time applications. *Proceedings of the IEEE*, 82(1), 25-40.
- [10] Schirru, R., & Pereira, C. M. (2004). A real-time artificially intelligent monitoring system for nuclear power plants operators support. *Real-Time Systems*, 27(1), 71-83.

- [11] Leucker, M., & Schallhart, C. (2009). A brief account of runtime verification. *The Journal of Logic and Algebraic Programming*, 78(5), 293-303.
- [12] Chodrow, S. E., Jahanian, F., & Donner, M. (1991, December). Run-time monitoring of real-time systems. In *Real-Time Systems Symposium, 1991. Proceedings., Twelfth* (pp. 74-83). IEEE.
- [13] Singer, R. M., Gross, K. C., Herzog, J. P., King, R. W., & Wegerich, S. (1997). Model-based nuclear power plant monitoring and fault detection: Theoretical foundations (No. ANL/RA/CP--91903; CONF-970765--1). Argonne National Lab., IL (United States).
- [14] Schreiber, F. A., Baiguera, M., Bortolotto, G., & Caglioti, V. (1997). A study of the dynamic behaviour of some workload allocation algorithms by means of catastrophe theory. *Journal of systems architecture*, 43(9), 605-624.
- [15] Van der Maas, H. L., & Molenaar, P. C. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological review*, 99(3), 395.
- [16] Van der Maas, H. L., Kolstein, R., & Van Der Pligt, J. (2003). Sudden transitions in attitudes. *Sociological Methods & Research*, 32(2), 125-152.
- [17] Sussmann, H. J., & Zahler, R. S. (1978). Catastrophe theory as applied to the social and biological sciences: A critique. *Synthese*, 37(2), 117-216.
- [18] Gilmore, R. (1993). Catastrophe theory for scientists and engineers. Courier Corporation.
- [19] Zeeman, E. C. (1977). Catastrophe theory: Selected papers, 1972–1977. Addison-Wesley.
- [20] Jerry, B. (1984). Discrete-event system simulation. Pearson Education India.
- [21] Nair, G. T., & Persya, C. A. (2012). Critical task re-assignment under hybrid scheduling approach in multiprocessor real-time systems. arXiv preprint arXiv:1203.6719.
- [22] Poston, T., & Stewart, I. (2014). Catastrophe theory and its applications. Courier Corporation.

Big Data Analytics Framework for Peer-To-Peer Botnet Detection Using Random Forest and Deep Learning

Saraladevi D. PG Student, *Department of Computer Science and Engineering, Pondicherry Engineering college.*

Sathiyamurthy K. Associate Professor, *Department of Computer Science and Engineering, Pondicherry Engineering College.*

Vijayaprabakaran K. PhD Scholar, *Department of Computer Science and Engineering, Pondicherry Engineering College.*

Abstract—Network traffic monitoring and analysis is constantly needed for protecting against botnet attacks. Existing solutions which usually rely on a high-performance server with huge storage capacity are not scalable for detailed analysis for large volume of traffic data. Detection and mitigation of peer-to-peer Botnet attacks in the distributed environment is the biggest challenge when gathering and analysing with big data. This paper explores how findings from machine learning with big data provide in higher prediction accuracy of malicious node in the network. An approach on Random forest is presented, but not deal with Overfitting reduces the prediction accuracy. In this work a Deep Neural network model increased the prediction accuracy and reduced computational complexity with H2O machine learning platform and integrated Big Data features. The goal of the implementation is used to detect Peer-to-Peer Botnet attacks using machine learning approach. The contributions of this paper are as follows: (1) Building a scalable framework for P2P botnet detection using Hadoop and H2O. (2) To adapt parallel processing power of H2O to build and compare Random Forest and Deep Learning implementation results.

Index Terms—Botnet, H2O, Big data, Peer-to-Peer, Machine Learning, Network Security, Deep Neural Network, Random Forest

I. INTRODUCTION

THIS Botnet attacks are one of the biggest challenges that security researchers and analysts face today on an

D. Saraladevi, PG Student, Department of Computer Science and Engineering, Pondicherry Engineering college, Puducherry, India. (saraladevi29@pec.edu).

K. Sathiyamurthy, Associate Professor, Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry, India (sathiyamurthyk@pec.edu)

K. Vijayaprabakaran, PhD Scholar, Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry, India (vijay.p.karan@gmail.com)

International scale. The economic losses were triggered by infringement of computer networks increases to billions of dollars. Just a few months prior to this writing, a massive DDoS attack had targeted the America's internet with a new weapon called Mirai botnet. The cause of the outage was a distributed denial of service (DDoS) attack, in which a network of computers infected with special malware, known as a "botnet", are coordinated to bombarding a server with traffic until it collapses under the strain. It was revealed that at least 90,000 unique IPs were utilized to achieve the attack [1]. It was doubtful that this attack was element of a bigger plan, where the attacker wanted to make an adjustment of numerous internet DNS infrastructure and then use the army of bots to launch an even larger DDoS attack. Hence, detecting and mitigating P2P Botnets and their attacks is still a challenge.

In order to detect and mitigate such attacks, network traces and packet captures are the most valuable resources for network security and network analysts. Such these attacks are growing exponentially day by day in the magnitude of network traces. However, computer systems lack the hardware and are fundamentally limited by the device. To address this issue of network security and threat detection, researchers have used various techniques such as Signature and Anomaly-based Intrusion Detection Systems (IDS). But these solutions have scalability issues when dealing with large datasets.

Though there were kernel scaling methods proposed [2], there were challenges with datasets having high variance. When there is larger dataset and has high variance, then it provides better the training accuracy of the model [3]. In the case of high variance, if data exceeds fitness level, then the training error will be low and then the cross-validation error will be much higher than the training error. It is assumed that if training is over-fitting the data; then the model will learn that data only but the model will generalize poorly for new observations and leading to a much higher cross-validation error. This can be corrected by increasing the size of the training data, that will reduce cross-validation error and in case of variance will be high.

A typical assumption of Big Data is that more data can lead to deeper vision and higher business value. This is especially true in machine learning, where algorithms can learn better from bigger data sets. However, massive data sets can be challenging to process [4, 5]. Many machine learning algorithms were designed with the perception that the entire dataset will fit into the memory [4]. Very often these algorithms are of high algorithmic complex in nature and needs huge volume of memory [6]. It leads to rise of various distributed processing approaches (i.e. MapReduce) which are appropriate for algorithms that can be parallelized to a degree sufficient to take advantage of available nodes.

The data sets demand the use of Big Data Technologies. Several large datasets have the malign activity of various bots have been captured and it is revealed by CAIDA and other authorized organizations. This research used traces captured by UCSD (University of California San Diego), of size 40 GB, [7] and it required the existence of a scalable framework to train the classification module.

In this paper a scalable distributed intrusion detection system has been proposed which can handle heavy network bandwidths. This framework is built on top of apache Spark, that is a free-source framework that provisions data-intensive distributed applications and influences the libraries which are built to use the supremacy of clustered commodity machines. The proposed framework makes use of H2O an open source artificial intelligence platform, which has machine learning and deep learning algorithms to build our predictive models.

The data analytics is done on the machine learning algorithms such as Random Forest and Multilayer Feed Forward Neural Network. Random forest is an ensemble model that can be used for both classification and regression. It operates by constructing a multitude of decision trees at training time; prediction then combines the outputs given by the individual trees to arrive at the final output. A key feature of random forests is that they correct for decision trees' tendency to overfit to their training data which reduces in prediction accuracy.

Training MLF-NN is a challenging task. It aims at learning features and hierarchies from lower level composition to higher levels. Multi-Layer Feed Forward network is implemented since botnet detection has the requirements of high accuracy of prediction and classification, ability to handle diverse bots, ability to handle data characterized by a very large number and diverse types of descriptors, ease of training, and computational efficiency.

The rest of the paper is organized as follows. Section 2 details the most relevant work in the domain of P2P Botnet detection which is based on machine learning algorithms and H2O applications in this area. Section 3 describes the experimental setup and the methodology utilized for the framework to accomplish security threat detection in real-life. The section 4 describes about the specific application of P2P Botnet detection using this framework and finally section 5 concludes with results and scope for further enhancement of work in future.

II. RELATED WORKS

Over the past few years, many research experts have proposed several solutions based on machine learning for mitigating security threats. Research has completely drifted from signature-based methods to more semantic-based methods for building Ontological models to manage web application attacks [8] and Hidden Markov Models (HMM) for spam identification [9]. DDoS attack detection has some patterns different with genuine traffic, it detects DDoS attacks based on specific characteristic features, used some fields in the IP header to calculate DDoS attacks features which was proposed by researcher Reyhaneh Karimazad and Ahmad Faraahi [10]. Security aware agent based systems were used to handle achieve security at runtime [11]. Also, several researches are going on for mitigating security threats using large network traces [12] where authors proposed a DDoS attack detection model using Hadoop framework. They had constructed a novel traffic monitoring system in a scalable manner that does NetFlow analysis on multi-terabytes of Internet traffic [13].

In the above work [13] a MapReduce algorithm with a new input format that is good enough in operating libpcap files in parallel however their approach hardcodes the features that can be extracted from the libpcap files and thereby the user does not provide the flexibility to decide on the feature set based on the problem instance. As per the current knowledge of the authors, the area of network security analytics severely lacks prior research in addressing the issue of Big Data.

On the other hand, there has been remarkable investigation in the domain of security threat detection via machine learning techniques. Authors [14] differentiate the network flow records depends on certain features related to traffic volume and classify them as malicious and benign. They also present how the plotters could change their behaviour to evade their detection technique, which was observed in Nugache, which is known to arbitrarily change its behavior. Authors in [15] show that P2P Botnet detection can be achieved with high accuracy based on a novel Bayesian Regularized Neural Network.

In [16] authors provide an overview on the adversarial attacks against IDSs and highlight the most promising research directions for the design of adversary-aware, harder to defeat IDS solutions.

Authors in [17] use P2P flow identification techniques to track and screen traffic flows, isolating the hosts once they connect to the Botnet. They deployed Bayes Classifier and Neural Network classifier in order to find the IP of the infected systems. Zhao et. al [18] in their study compared two machine learning techniques, decision tree and Naive Bayes for P2P botnet detection. The experiments were conducted using traces of three botnets: Storm, Nugache and Waledac. A powerful command-line packet analyser, Tcp-dump is used for network traffic capture [19]. Other visually rich network protocol analyser called Wireshark is used in [20], which more powerful LAN analyser tool that captures live traffic and stores it for offline analysis.

Authors in [21] applied a self-organization map algorithm to determine P2P Botnets, in that they assume there would be several failed connection accomplishments from exterior to interior in firewall. And in [22] authors have presented a model based on Discriminative Restricted Boltzmann Machine to combine the expressive power of generative models with good classification accuracy to infer knowledge from incomplete training data.

Zhang et al. [23] reviewed in memory Big Data management and processing. They distinguished in-memory systems as batch-oriented systems such as Spark and H2O, and real time or stream processing systems such as Storm. The systems relevant to botnet prediction primarily belong to the batch category.

The publications of Chen and Lin [24] and Najafabadi et al. [25] examined deep learning with Big Data and discussed the associated challenges. Both studies highlighted the role of dimensionality reduction, parallel processing, and distributed processing in deep network training. Our work takes advantage of parallel and distributed processing and performs dimensionality reduction, but only after the training data have been partitioned into clusters. Al-Jarrah et al. [26] reviewed energy efficient machine learning approaches and new approaches with reduced memory requirements. They thought local learning as one of the key mechanisms for machine learning techniques with Big Data due to its computation cost. They considered deep learning to be a key technique to provide representation learning for complex problems. H2O deep learning is an example of recent deep learning approaches for Big Data.

Most among the previous work, researchers have concentrated on detecting a particular Botnet activity and their methods were not reported to be successful in detecting bots, whose traffic characteristics were not used in the training set. Clearly, it can be seen that using a machine learning based approach is far superior in detecting malicious traffic when compared to a traditional signature based approach as the bot masters redesign the bots from time to time and the functionality and behavior of the Botnet varies quite significantly with each version release of the bot. And signature based detectors rely heavily upon the existing Botnet signatures to detect any activity.

Particularly when handling zero-day attacks, signature based approach fails completely as there is no account of previous activity for that bot. Therefore, a machine learning approach is usually preferred to find seemingly suspicious activity based on the anomalous behaviour of the network. In the previous work, there was lack of research on deploying the detection module in real-world case to monitor and reduce the Botnet activity in a network. In this work, this aspect of handling large-scale network traffic in a very short time is addressed and a solution is proposed to deal with Botnet detection at quasi-real time in heavy bandwidths of data traffic.

III. SCALABLE FRAMEWORK FOR P2P BOTNET DETECTION

The proposed framework includes the technologies such as Libpcap, spark, MapReduce and H2O. To extract the desired

fields from the packets Tshark was used. These technologies are need to generate the feature set for the Machine Learning Module. Libpcap library used in the Tshark, that captures packet data from a live network. Also, it allows printing a decoded and customizable form of the captured packets to the standard output or a file. After the extraction of desired information from the Sniffer Module, the feature extraction was done by MapReduce. This can be achieved using Apache Hive which provides a mechanism to query the data using HiveQL (a SQL-like language).

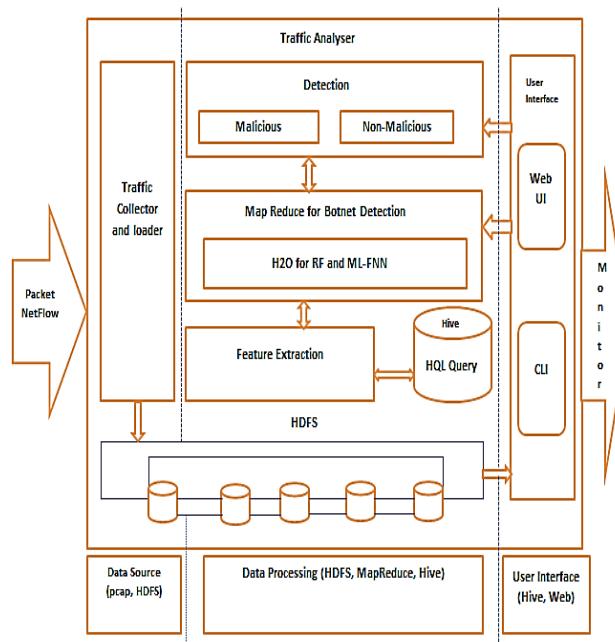


Figure.1 Scalable Framework for Botnet Detection

The Fig.1 shows the Scalable Framework for Botnet Detection. The process starts with pre-processing packets from the network interface and then features re extracted using Tshark. The extracted features are loaded into H2O for classifying the data as malicious and non-malicious using random forest and deep learning approach.

This section discussed about the components of the proposed scalable network threat detection framework in detail. The proposed framework consists of the following components:

- Traffic Sniffer Module for packet pre-processing packets.
- Feature Extraction Module for generating feature set.
- Machine Learning Module for learning and detecting the malicious traffic.

These blocks are detailed and discussed in the subsequent three subsections, highlighting the problems faced and solutions proposed in the implementation.

A. Traffic Sniffer

Dumpcap [27] is used for sniffing the packets from the network interface while Tshark [28] is used for extracting the fields, related to feature set, out of the packets and to submit the fields to the HDFS. Even though Tshark could have been

used for sniffing the packets from the interface, dumpcap gives better performance when dealing with long term captures [29] since it is the lowest level abstraction of libpcap. Whereas the constrained buffers of Tshark present more overhead and consume more time. Thus, dumpcap capture ring buffer option is used to capture traffic onto successive pcaps which are then handled by multiple Tshark instances thereby achieving a greater degree of parallelism.

The traffic Sniffer Module saves the traffic from the wire into successive pcap files of a specified size using the capture ring buffer option.

The delimited files from each of these Tshark instances which run in parallel are all submitted to the HDFS upon completion. These are then loaded into a table via the “LOAD DATA” command in HQL. The samples of these malware were obtained from [30, 31].

Three different botnet dataset each consists of around 20,000 instances are collected. The normal dataset is captured in real-time using wireshark tool.

The network activity of these malware samples was monitored and captured by Wireshark. After collecting packet traces of the network activity of each of the malware, the pcaps were saved in the H2O cluster for further investigations. As executable files for all bots were not available, few of the already captured packets from web communities like CAIDA, ContagioDump were used.

An optimal file size of 2 GB was opted for the pcap to perform proper classification, benign traffic involves traffic from P2P application, ftp transfers, telnet sessions, video streaming and mobile updates were collected by the authors. Thus, the entire dataset was aggregated and given to the machine learning algorithm for model generation.

The steps followed for data Pre-processing are

- Data Cleaning: process of determining and removing errors, filling missing values of data and inconsistencies from data in order to enhance the quality.
- Data Integration: malicious and non- malicious datasets are integrated.
- Normalization: IP addresses are converted to unique numerical values. For example, the normalization function converts the IP address 143.31.1.101 as 100 and this value will be assigned for every occurrence of this IP address.
- Duplicate Elimination: editcap program is used to remove duplicate packets exits in the traffic. This program compares the current entry with the previous ‘n’ packets in the log in which n specifies the window size and eliminates this current entry if it exists in the ‘n’ packets. Pass final instance set to Botnet Classifier.

B. Traffic Sniffer

Once the delimited files are submitted to HDFS, Apache Hive [32] is used to extract the features out of them. One of the important features in this framework is the ability to change the features at runtime which is enabled by Apache Hive and Tshark. The Feature Extraction Perl Script enables the user to decide the fields to extract from packet using Tshark and then creates the table in Hive accordingly. In case a different feature set is chosen in this script, a different table is created automatically.

The Apache Hive provides extract/transform/load (ETL) and managing large datasets which is built over the Apache spark. It provides an easy understandable SQL like language called HQL [33]. Hive translates these Hive QL into MapReduce programs and executes in runtime.

Table 1. Flow statistics that can be extracted from the software

Feature	Description of Feature
Srcip	Source ip address (string)
Srcport	Source port number
Dstip	Destination ip address (string)
Dstport	Destination port number
Proto	Protocol (ie. TCP =6, UDP=17)
Total_fpackets	Total packets in forward direction
Total_fvolume	Total bytes in forward direction
Total_bpackets	Total packets in backward direction
Min_fpktl	Size of smallest packet sent in forward direction (in bytes)
Mean_fpktl	Mean size of packets sent in forward direction (in bytes)
Max_fpktl	Size of largest packet sent in forward direction (in bytes)
Std_fpktl	Standard deviation from mean of packets sent in forward direction (in bytes)
Min_bpktl	Size of smallest packet sent in backward direction (in bytes)
Mean_bpktl	Mean size of packets sent in backward direction (in bytes)
Max_bpktl	Size of largest packet sent in backward direction (in bytes)
Std_bpktl	Standard deviation from mean of packets sent in backward direction (in bytes)
Min_fiat	Minimum amount of time between two packets sent in forward direction (in microseconds)
Mean_fiat	Mean amount of time between two packets sent in forward direction (in microseconds)
Max_fiat	Maximum amount of time between two packets sent in forward direction (in microseconds)
Std_fiat	Standard deviation from mean amount of time between two packets sent in forward direction (in microseconds)
Min_biat	Minimum amount of time between two packets sent in backward direction (in microseconds)

Mean_biat	Mean amount of time between two packets sent in backward direction (in microseconds)
Max_biat	Maximum amount of time between two packets sent in backward direction (in microseconds)
Duartion	Duration of flow (in microseconds)
Total_bhflen	Total bytes used for headers in backward direction
Total_fhflen	Total bytes used for headers in forward direction

Since most of the features extracted for this problem, which are shown in Table 1, are flow based statistics such as size of largest packet in a flow, they are extracted out of the table using the “group by” clause in HQL. The group by is based on MapReduce algorithm. The map phase generates the key-value pairs which are passed onto the reduce phase. Here the reducer groups all the values based on the key passed to it. That is, the MapReduce framework operates exclusively on $\langle \text{key}, \text{value} \rangle$ pairs where the input to framework is a set of $\langle \text{key}, \text{value} \rangle$ pairs and the output produced by the job is also another set of $\langle \text{key}, \text{value} \rangle$ pair.

(Input) $\langle k1, v1 \rangle \rightarrow \text{map} \rightarrow \langle k2, v2 \rangle \rightarrow \text{combine} \rightarrow \langle k2, v2 \rangle \rightarrow \text{reduce} \rightarrow \langle k3, v3 \rangle$ (output)

Here, the data is grouped by source IP, source port, destination IP and destination port to extract flows from raw packet data. So, the key here is a combination of source IP, source port, destination IP, destination port. The values here are the remaining fields which were a part of the delimited files generated by Tshark. The mapper of H2O generates the key value pairs which are then passed to reducer which groups all the packets for that particular key flow and generates the features like total bytes transferred or average inter arrival time in a flow.

Table 2. Features Selected Using Information Gain Ranking Algorithm

Rank	Feature
0.851	total_bhflen
0.823	Std_bpktl
0.803	fpsh_cnt
0.7956	total_fhflen
0.7886	bpsh_cnt
0.7438	min_biat
0.7182	min_fiat

Then Information Gain Attribute Evaluation was done using the Ranker Algorithm in order to find out the most influential features of the entire feature set. This method evaluates the worth of an attribute by measuring the Information Gain with respect to the class, where Information Gain is described by the following equation:

Information Gain (Class, Attribute) - H (Class) – H (Class| Attribute).

The feature set used in this context along with the corresponding Info Gain are presented in Table 2, which describes the priority of features.

C. Feature Extraction Module

H2O - a machine learning library built on top of spark was used to achieve scalability in Machine Learning. Since each of its core algorithms for classification and clustering are run in parallel as MapReduce jobs, the high computational power of the cluster is harnessed to attain optimized results [35].

Implementing a non-distributed classifier leads to big load for a single system to deal with since the data alone could consume the entire heap space of JVM on top of which most of the APIs of the classifiers run such as Weka [36]. Other existing implementations (C/C++) of classifiers which executes on standalone systems needs lot of resources and often run out of memory when dealing with Big Data. Hence, the distributed deployment of H2O plays a major role in Big Data Analysis. H2O native library of Random Forest's and Multi-Layer Feed Forward Neural Network implementation is used for model training and classification purpose.

First a file descriptor is created for the dataset which is the set of features mined from previous module which consists of various types of attributes such as numeric, label and categorical. Then Hundred trees with 5 arbitrarily selected attributes per node were built. And then the predicted outcome for each of the test instances are stored onto the HDFS, that can be accessed using Hue Web UI and malicious nodes can be determined respectively in random forest.

- Naive implementation of random forests on distributed systems easily overfits the training data, yielding poor classification performances for Peer-to-peer botnet detection and also Not good for regression problem as it does not give precise continuous nature predictions. This is avoided by the deep learning since it focuses on regularization techniques such as lasso and ridge.
- Multi-layer feed forward neural network (MLFNN) contains many layers of interconnected neuron units, beginning with an input layer to match the feature space followed by multiple layers of nonlinearity and ending with a linear regression or classification layer to match the output space. Bias units are incorporated in each non-output layer of the network. The weights connecting neurons and biases with other neurons fully determine the output of the entire network, and learning occurs when these weights are adjusted to limit the error on named training data of the botnet.
- For each training data j , the objective is to limit a loss function,

$$L(W, B | j).$$

Where

- W is the collection $\{W_i\}_{1:N-1}$, where W_i represents the weight matrix of connecting layers i and $i + 1$ for a network of N layers.
- B is the collection $\{b_i\}_{1:N-1}$, where b_i denotes the column vector of biases for layer $i + 1$.

H2O framework for multi-layer neural networks architectures are models of hierarchical feature extraction, it involves multiple levels of nonlinearity.

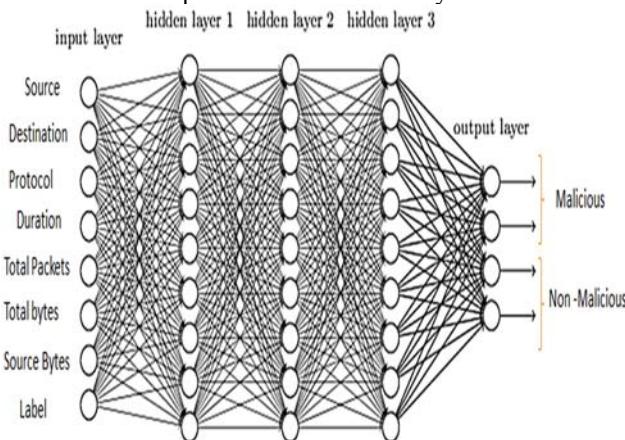


Figure.2 Multi-Layer Feed Forward Neural Network

For fully connected networks, number of parameters (N) is given in eqn.1

$$N = d \times N^{(l)} + K \times N^{(L)} + \sum_{k=1}^{L-1} N^k \times N^{k+1} . \quad (\text{eqn.1})$$

- L is the number of hidden layers where the packet features to be added.
- $N(k)$ is the number of nodes for layer k .
- d is the input vector size of MLFNN matrices.
- K is the output size which describes benign and malicious.

To classify the malicious and non – malicious traffic in the network, the features are extracted with different hidden layers in the above eqn.1 and hence node values are calculated.

1) Parallel distributed and multi-threaded training in H2O Deep Learning

Stochastic gradient descent is well known for being fast and memory efficient, but not simply parallelizable unless becoming slow. To address this issue, Hogwild! approach is used which follows a shared memory model where multiple cores of each handling separate subsets or all of the training data. This makes unconstrained contributions to the gradient updates $\nabla L(W, B | j)$ asynchronously. In a multi-node system, this parallelization scheme runs on top of H2O's distributed setup and distributes the training data over the cluster.

2) Parallelization Scheme on H2O

1. First, Initialize the global model parameters W, B
2. Distribute training data T across nodes.
3. Repeat until convergence condition reached:
 - 3.1 For n nodes with training subset T_n , do in parallel:
 - a. Get replica of the global model parameters W_n, B_n

- b. Select active subset $T_{na} \subset T_n$ (user-given number of samples per iteration)
- c. Separate T_{na} into T_{nac} by cores n_c
- d. For cores n_c on node n , do in parallel:
 - i. Obtain training example $i \in T_{nac}$
 - ii. Update all weights $w_{jk} \in W_n$, biases $b_{jk} \in B_n$ $w_{jk} := w_{jk} - \alpha \partial L(W, B) / \partial w_{jk}$
 $b_{jk} := b_{jk} - \alpha \partial L(W, B) / \partial b_{jk}$
- 3.2 Assign $W, B := \text{Avg}_n W_n, \text{Avg}_n B_n$
- 3.3 Optionally score the model on potentially sampled train/validation scoring gets.

In this algorithm, each node operates in parallel on its local data until the final parameters W, B are obtained by averaging. Here, the weights and bias updates follow the asynchronous Hogwild! Procedure to incrementally adjust each node's parameters W_n, B_n . The Avg notation refers to the final averaging of these local parameters across all nodes to obtain the global model parameters and complete training.

D. Algorithm for Multi-Layer Feed Forward Neural Network

1. Train the first layer as an MLFNN that models the raw input $X=h(0)$ as its visible layer.
2. Take that first layer to get a representation of the input that will become input data for the second layer. Two common solutions exist. This representation can be selected as being the mean activations $p(h(1) = 1|h(0))$ or samples of $p(h(1)|h(0))$.
3. Then, train the second layer as an FNN, taking the transfigured data samples or mean activations as training instances for the visible layer of that DNN.
4. Repeat 2 and 3 for the desired number of layers, it grows upward either samples or mean values in iteration.
5. Adjust all the parameters of this deep architecture with respect to response column, activation, hidden layers, epochs, variable importance, adaptive rate, hidden dropout ratios, $l1$ and $l2$ regularization.

Models based on k-Nearest Neighbors, ANN and nonlinear SVMs have very high prediction rates and are versatile in taking into consideration the variety of the training data while constructing the model. However, ANN and k-NN not so efficient when there is high-dimensional data without dimension reduction or pre-selection of descriptors. While Nonlinear SVM is robust to the presence of a large number of irrelevant descriptors, thus necessitating descriptor preselection. Random Forest does not yield with overfitting of data. Several attempts have been made to addressing this problem using Independent Component Analysis as in [37] yet it might not be effective always to handle lower dimensional data.

Thus, Deep Learning is the closest to have the craved combination of features. It is great at dealing high-dimensional data and disregards inappropriate descriptors by pruning them. One such algorithm improvised to Multi-Layer Feed Forward Neural Network.

IV. RESULTS

The classification modules are trained with capture files from well-known Bot attacks (e.g. Kelihos-Hlux, Conficker, Storm, Zeus, and Waledac. These datasets of the bot attack with some benign traffic are captured in PCAP format, that is also required in classifier.

A total of 2,84,030 examples of varied traffic was used as the dataset from which 80% was taken as the training set and 20% was taken as testing set. The classifier provided an accuracy of 90.3% when deployed by means of Random Forest Algorithm where 10 trees were used to build the forest. In case of ML-FNN, the classifier gave an accuracy of 98.41. This shows the deep learning provides much higher level of prediction of accuracy as compared with other machine learning models.

The Table 3 shows the precision, recall and F- score of the Random Forest and the Multi-Layer Feed Forward Neural Network for the different botnet data data size with heterogeneous data.

Table 3. Table for Evaluation Metrics

Size of dataset <i>t</i>	Random Forest			Multi-Layer Feed Forward Neural Network		
	Precisio n	Recall	F-Score	Precisio n	Recall	F-Score
20k	0.9994	0.841	0.913	0.9873	0.975	0.987
		0	4		1	4
40k	0.9996	0.804	0.891	0.9440	1.0	0.971
		8	7			2
60k	0.9996	0.799	0.888	1.0	0.919	0.957
		1	2		2	9
80k	0.9996	0.782	0.878	1.0	0.900	0.947
		8	0		8	8
1lk	0.9996	0.739	0.850	1.0	0.880	0.936
		6	1		4	4
1.5l	0.9995	0.689	0.816	1.0	0.858	0.924
		4	0		9	1
2l	0.9994	0.580	0.734	1.0	0.838	0.912
		9	7		9	1

Table 4. Table for Accuracy

Size of dataset	RF Accuracy (%)	ML-FNN Accuracy (%)
40k	90.30	98.41
80k	88.11	98.30
1l	87.66	97.06
2l	86.77	96.23
3l	84.14	95.89
5l	81.09	94.88
10l	74.49	94.73

The Table 4 shows the accuracy of the proposed system and existing system for the different data size. The random forest approach for detecting peer to peer botnet leads to decrease in accuracy as the trees grows larger. This tends to the complexity of the data and it is avoided by the multi-layer feed forward neural network.

The Figure 3 illustrates the comparison of accuracy between the Random Forest algorithm and Multi-layer feed forward neural network algorithm. The graph shows the higher prediction efficiency in ML-FNN algorithm when compared to Random Forest algorithm since ML-FNN deals with overfitting and complexity of the data. The parallel and distributed process is achieved by apache spark and H2O which handles big data.

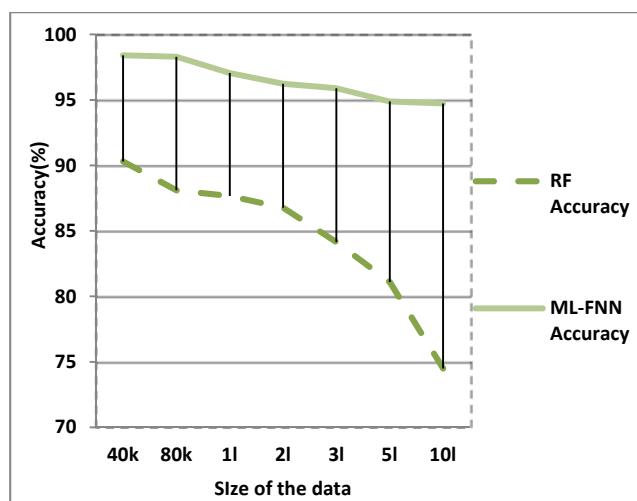


Figure.3 Comparison of accuracy

V. SOME COMMON MISTAKES

This research paper contributes in the following modules:

1. A scalable packet capture module to process large bandwidths of data in a quasi-real-time.
2. A distributed dynamic feature mining framework to describe flow statistics of packet captures.
3. A Peer-to-Peer security threat detection module of

Random Forest and ML-FNN which classifies malicious traffic on a cluster.

Detection of botnet attack in the peer to peer network has been identified using machine learning approach with big data analytics. Accuracy of finding the malicious node attained maximum level using multi-layer feed forward neural network. The future work can be extended with providing security recommendations such as firewall implementation for the malicious node to avoid further attacking the other nodes in the network. This approach can be further used for finding other network attacks such as man in the middle attack, backdoor attacks with higher classification prediction accuracy.

VI. CONCLUSION AND FUTURE WORK

A. Types of Graphics

The This research paper contributes in the following modules:

1. A scalable packet capture module to process large bandwidths of data in a quasi-real-time.
2. A distributed dynamic feature mining framework to describe flow statistics of packet captures.
3. A Peer-to-Peer security threat detection module of Random Forest and ML-FNN which classifies malicious traffic on a cluster.

Detection of botnet attack in the peer to peer network has been identified using machine learning approach with big data analytics. Accuracy of finding the malicious node attained maximum level using multi-layer feed forward neural network. The future work can be extended with providing security recommendations such as firewall implementation for the malicious node to avoid further attacking the other nodes in the network. This approach can be further used for finding other network attacks such as man in the middle attack, backdoor attacks with higher classification prediction accuracy.

REFERENCES

- [1] <https://www.theguardian.com/technology/2016/oct/26/ddos-attack-dyn-mirai-botnet>
- [2] [A. Maratea, A. Petrosino, M. Manzo, Adjusted F-measure and kernel scaling for imbalanced data learning, Inf. Sci. 257 \(2014\) 331–341.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4117033/)
- [3] Ng Andrew, Advice for Applying Machine Learning, 2013.
- [4] K. Grolinger, M. Hayes, W. A. Higashino, A. L'Heureux, D. S. Allison, and M. A. M. Capretz, "Challenges for MapReduce in Big Data," *Proceedings of the IEEE World Congress on Services*, pp. 182– 189, 2014.
- [5] K. Grolinger, W. A. Higashino, A. Tiwari, and M. A. Capretz, "Data management in cloud environments: NoSQL and NewSQL data stores," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 2, 2013.
- [6] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidatc, G. K. Karagiannidis, and K. Tahaa, "Efficient machine learning for Big Data: A review," *Big Data Research*, vol. 2, no. 3, pp. 87–93, 2015.
- [7] The CAIDA UCSD Dataset 2008-11-21,2008.<<https://data.caida.org/datasets/security/telescope-3days-conficker/>>.
- [8] [A. Razzaq, K. Latif, H.F. Ahmad, A. Hur, Z. Anwar, P.C. Bloodsworth, Semantic security against web application attacks, Inf. Sci. 254 \(2014\) 19–38.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3990712/)
- [9] [F. Salcedo-Campos, J. Diaz-Verdejo, P. Garcia-Teodoro, Segmental parameterisation and statistical modelling of e-mail headers for spam detection, Inf](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3990712/)
- [10] Reyhaneh Karimazad and Ahmad Faraahi. An Anomaly-Based Method for DDoS Attacks Detection using RBF Neural Networks. *IPCSIT* vol.11 (2011) © (2011) IACSIT Press, Singapore.
- [11] [G. Beydoun, G. Low, Generic modelling of security awareness in agent based systems, Inf. Sci. 239 \(2013\) 62–71.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3990712/)
- [12] [Y. Lee, Y. Lee, Toward scalable internet traffic measurement and analysis with hadoop, ACM SIGCOMM Comput. Commun. Rev. 43 \(1\) \(2012\) 5–13.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3990712/)
- [13] Analysis, Springer, Berlin Heidelberg, 2011, pp. 51–63.
- [14] T.F. Yen, M.K. Reiter, Are your hosts trading or plotting? Telling P2P file-sharing and bots apart, in: *2010 IEEE 30th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, June 2010, pp. 241–252.
- [15] S.C. Guntuku, P. Narang, C. Hota, Real-Time Peer-to-Peer Botnet Detection Framework based on Bayesian Regularized Neural Network, 2013, arXiv preprint arXiv:1307.7464.
- [16] [I. Corona, G. Giacinto, F. Roli, Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues, Inf. Sci. 239 \(2013\) 201– 225.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3990712/)
- [17] [W. Tarn, L.Z. Den, K.L. Ou, M. Chen, The analysis and identification of P2P botnet's traffic flows, Int. J. Commun. Netw. Inf. Secur. \(IJCNIS\) 3 \(2\) \(2011\).](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3990712/)
- [18] D. Zhao, I. Traore, A. Ghorbani, B. Sayed, S. Saad, and W. Lu, "Peer to peer botnet detection based on flow intervals," in *Information Security and Privacy Research*. Springer, 2012, pp. 87–102.
- [19] Tcpdump, <http://www.tcpdump.org>
- [20] Wireshark, <http://www.wireshark.org>
- [21] C. Langin, H. Zhou, S. Rahimi, B. Gupta, M. Zargham, M.R. Sayeh, A self-organizing map and its modeling for discovering malignant network traffic, in: *IEEE Symposium on Computational Intelligence in Cyber Security*, 2009, CICS'09, IEEE, March 2009, pp. 122–129.
- [22] [U. Fiore, F. Palmieri, A. Castiglione, A. De Santis, Network anomaly detection with the restricted Boltzmann machine, Neurocomputing 122 \(2013\) 13– 23.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3990712/)
- [23] H. Zhang, G. Chen, and B. Ooi, "In-memory Big Data management and processing: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 7, pp. 1920–1948, 2015.
- [24] X.-W. Chen and X. Lin, "Big Data deep learning: challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.
- [25] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, pp. 1–21, 2015.
- [26] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidatc, G. K. Karagiannidis, and K. Tahaa, "Efficient machine learning for Big Data: A review," *Big Data Research*, vol. 2, no. 3, pp. 87–93, 2015.
- [27] Dumpcap, 2016. <<http://www.wireshark.org/docs/man-pages/dumpcap.html>>.
- [28] TShark, 2016. <<http://www.wireshark.org/docs/man-pages/tshark.html>>.
- [29] Network Protocol Specialists, 2008. <<http://new.networkprotocolspecialists.com/tips/long-term-captures-with-tshark/>>
- [30] ContagioDumpBlogspot,2013.<<http://contagiodump.blogspot.in/>>.
- [31] Open Malware, 2013. <<http://openmalware.org/>>.
- [32] Apache Hive, 2016. <hive.apache.org>.
- [33] A. Thusoo, J.S. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, ... R. Murthy, Hive-a petabyte scale data warehouse using hadoop, in: *2010 IEEE 26th International Conference on Data Engineering (ICDE)*, IEEE, March 2010, pp. 996–1005.
- [34] D.Arndt,Netmate,2015.<<http://dan.arndt.ca/nims/calculating-flow-statistics-using-netmate>>
- [35] H2O ,2016 .<<https://www.h2o.ai/sparkling-water/>>.

- [36] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten,
The WEKA data mining software: an update, *ACM SIGKDD Explor.*
Newslett. 11 (1) (2009) 10–18.
- [37] F. Palmieri, U. Fiore, A. Castiglione, A distributed approach to network
anomaly detection based on independent component analysis, *Concurr.*
Comput.:

IJCSIS REVIEWERS' LIST

- Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore, Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Dr. Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Dr Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Dr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Dr. Somesh Kumar Dewangan, CCSVTU Bhilai (C.G.)/ Dimat Raipur, India
Dr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Dr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Dr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Dr. P. Vasant, University Technology Petronas, Malaysia
Dr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Dr. Praveen Ranjan Srivastava, BITS PILANI, India
Dr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Dr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Dr. Tirthankar Gayen, IIT Kharagpur, India
Dr. Huei-Ru Tseng, National Chiao Tung University, Taiwan
Prof. Ning Xu, Wuhan University of Technology, China
Dr Mohammed Salem Binwahlan, Hadhramout University of Science and Technology, Yemen
& Universiti Teknologi Malaysia, Malaysia.
Dr. Aruna Ranganath, Bhoj Reddy Engineering College for Women, India
Dr. Hafeezullah Amin, Institute of Information Technology, KUST, Kohat, Pakistan

Prof. Syed S. Rizvi, University of Bridgeport, USA
Dr. Shahbaz Pervez Chattha, University of Engineering and Technology Taxila, Pakistan
Dr. Shishir Kumar, Jaypee University of Information Technology, Wakanaghata (HP), India
Dr. Shahid Mumtaz, Portugal Telecommunication, Instituto de Telecomunicações (IT) , Aveiro, Portugal
Dr. Rajesh K Shukla, Corporate Institute of Science & Technology Bhopal M P
Dr. Poonam Garg, Institute of Management Technology, India
Dr. S. Mehta, Inha University, Korea
Dr. Dilip Kumar S.M, Bangalore University, Bangalore
Prof. Malik Sikander Hayat Khiyal, Fatima Jinnah Women University, Rawalpindi, Pakistan
Dr. Virendra Gomase , Department of Bioinformatics, Padmashree Dr. D.Y. Patil University
Dr. Irraivan Elamvazuthi, University Technology PETRONAS, Malaysia
Dr. Saqib Saeed, University of Siegen, Germany
Dr. Pavan Kumar Gorakavi, IPMA-USA [YC]
Dr. Ahmed Nabih Zaki Rashed, Menoufia University, Egypt
Prof. Shishir K. Shandilya, Rukmani Devi Institute of Science & Technology, India
Dr. J. Komala Lakshmi, SNR Sons College, Computer Science, India
Dr. Muhammad Sohail, KUST, Pakistan
Dr. Manjaiah D.H, Mangalore University, India
Dr. S Santhosh Baboo, D.G.Vaishnav College, Chennai, India
Prof. Dr. Mokhtar Beldjehem, Sainte-Anne University, Halifax, NS, Canada
Dr. Deepak Laxmi Narasimha, University of Malaya, Malaysia
Prof. Dr. Arunkumar Thangavelu, Vellore Institute Of Technology, India
Dr. M. Azath, Anna University, India
Dr. Md. Rabiul Islam, Rajshahi University of Engineering & Technology (RUET), Bangladesh
Dr. Aos Alaa Zaidan Ansaef, Multimedia University, Malaysia
Dr. Suresh Jain, Devi Ahilya University, Indore (MP) India,
Dr. Mohammed M. Kadhum, Universiti Utara Malaysia
Dr. Hanumanthappa. J. University of Mysore, India
Dr. Syed Ishtiaque Ahmed, Bangladesh University of Engineering and Technology (BUET)
Dr Akinola Solomon Olalekan, University of Ibadan, Ibadan, Nigeria
Dr. Santosh K. Pandey, The Institute of Chartered Accountants of India
Dr. P. Vasant, Power Control Optimization, Malaysia
Dr. Petr Ivankov, Automatika - S, Russian Federation
Dr. Utkarsh Seetha, Data Infosys Limited, India
Mrs. Priti Maheshwary, Maulana Azad National Institute of Technology, Bhopal
Dr. (Mrs) Padmavathi Ganapathi, Avinashilingam University for Women, Coimbatore
Assist. Prof. A. Neela madheswari, Anna university, India
Prof. Ganesan Ramachandra Rao, PSG College of Arts and Science, India
Mr. Kamanashis Biswas, Daffodil International University, Bangladesh
Dr. Atul Gonsai, Saurashtra University, Gujarat, India
Mr. Angkoon Phinyomark, Prince of Songkla University, Thailand
Mrs. G. Nalini Priya, Anna University, Chennai
Dr. P. Subashini, Avinashilingam University for Women, India
Assoc. Prof. Vijay Kumar Chakka, Dhirubhai Ambani IICT, Gandhinagar ,Gujarat
Mr Jitendra Agrawal, : Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal
Mr. Vishal Goyal, Department of Computer Science, Punjabi University, India
Dr. R. Baskaran, Department of Computer Science and Engineering, Anna University, Chennai

Assist. Prof, Kanwalvir Singh Dhindsa, B.B.S.B.Engg. College, Fatehgarh Sahib (Punjab), India
Dr. Jamal Ahmad Dargham, School of Engineering and Information Technology, Universiti Malaysia Sabah
Mr. Nitin Bhatia, DAV College, India
Dr. Dhavachelvan Ponnurangam, Pondicherry Central University, India
Dr. Mohd Faizal Abdollah, University of Technical Malaysia, Malaysia
Assist. Prof. Sonal Chawla, Panjab University, India
Dr. Abdul Wahid, AKG Engg. College, Ghaziabad, India
Mr. Arash Habibi Lashkari, University of Malaya (UM), Malaysia
Mr. Md. Rajibul Islam, Ibnu Sina Institute, University Technology Malaysia
Professor Dr. Sabu M. Thampi, .B.S Institute of Technology for Women, Kerala University, India
Mr. Noor Muhammed Nayeem, Université Lumière Lyon 2, 69007 Lyon, France
Dr. Himanshu Aggarwal, Department of Computer Engineering, Punjabi University, India
Prof R. Naidoo, Dept of Mathematics/Center for Advanced Computer Modelling, Durban University of Technology, Durban, South Africa
Prof. Mydhili K Nair, Visweswaraiah Technological University, Bangalore, India
M. Prabu, Adhiyamaan College of Engineering/Anna University, India
Mr. Swakkhar Shatabda, United International University, Bangladesh
Dr. Abdur Rashid Khan, ICIT, Gomal University, Dera Ismail Khan, Pakistan
Mr. H. Abdul Shabeer, I-Nautix Technologies, Chennai, India
Dr. M. Aramudhan, Perunthalaivar Kamarajar Institute of Engineering and Technology, India
Dr. M. P. Thapliyal, Department of Computer Science, HNB Garhwal University (Central University), India
Dr. Shahaboddin Shamshirband, Islamic Azad University, Iran
Mr. Zeashan Hameed Khan, Université de Grenoble, France
Prof. Anil K Ahlawat, Ajay Kumar Garg Engineering College, Ghaziabad, UP Technical University, Lucknow
Mr. Longe Olumide Babatope, University Of Ibadan, Nigeria
Associate Prof. Raman Maini, University College of Engineering, Punjabi University, India
Dr. Maslin Masrom, University Technology Malaysia, Malaysia
Sudipta Chattopadhyay, Jadavpur University, Kolkata, India
Dr. Dang Tuan NGUYEN, University of Information Technology, Vietnam National University - Ho Chi Minh City
Dr. Mary Lourde R., BITS-PILANI Dubai, UAE
Dr. Abdul Aziz, University of Central Punjab, Pakistan
Mr. Karan Singh, Gautam Budtha University, India
Mr. Avinash Pokhriyal, Uttar Pradesh Technical University, Lucknow, India
Associate Prof Dr Zuraini Ismail, University Technology Malaysia, Malaysia
Assistant Prof. Yasser M. Alginahi, Taibah University, Madinah Munawwarah, KSA
Mr. Dakshina Ranjan Kisku, West Bengal University of Technology, India
Mr. Raman Kumar, Dr B R Ambedkar National Institute of Technology, Jalandhar, Punjab, India
Associate Prof. Samir B. Patel, Institute of Technology, Nirma University, India
Dr. M. Munir Ahmed Rabbani, B. S. Abdur Rahman University, India
Asst. Prof. Koushik Majumder, West Bengal University of Technology, India
Dr. Alex Pappachen James, Queensland Micro-nanotechnology center, Griffith University, Australia
Assistant Prof. S. Hariharan, B.S. Abdur Rahman University, India
Asst Prof. Jasmine. K. S, R.V. College of Engineering, India
Mr Naushad Ali Mamode Khan, Ministry of Education and Human Resources, Mauritius
Prof. Mahesh Goyani, G H Patel College of Engg. & Tech, V.V.N. Anand, Gujarat, India
Dr. Mana Mohammed, University of Tlemcen, Algeria
Prof. Jatinder Singh, Universal Institution of Engg. & Tech. CHD, India

Mrs. M. Anandhavalli Gauthaman, Sikkim Manipal Institute of Technology, Majitar, East Sikkim
Dr. Bin Guo, Institute Telecom SudParis, France
Mrs. Maleika Mehr Nigar Mohamed Heenaye-Mamode Khan, University of Mauritius
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Mr. V. Bala Dhandayuthapani, Mekelle University, Ethiopia
Dr. Irfan Syamsuddin, State Polytechnic of Ujung Pandang, Indonesia
Mr. Kavi Kumar Khedo, University of Mauritius, Mauritius
Mr. Ravi Chandiran, Zagro Singapore Pte Ltd. Singapore
Mr. Milindkumar V. Sarode, Jawaharlal Darda Institute of Engineering and Technology, India
Dr. Shamimul Qamar, KSJ Institute of Engineering & Technology, India
Dr. C. Arun, Anna University, India
Assist. Prof. M.N.Birje, Basaveshwar Engineering College, India
Prof. Hamid Reza Naji, Department of Computer Enigneering, Shahid Beheshti University, Tehran, Iran
Assist. Prof. Debasis Giri, Department of Computer Science and Engineering, Haldia Institute of Technology
Subhabrata Barman, Haldia Institute of Technology, West Bengal
Mr. M. I. Lali, COMSATS Institute of Information Technology, Islamabad, Pakistan
Dr. Feroz Khan, Central Institute of Medicinal and Aromatic Plants, Lucknow, India
Mr. R. Nagendran, Institute of Technology, Coimbatore, Tamilnadu, India
Mr. Amnach Khawne, King Mongkut's Institute of Technology Ladkrabang, Ladkrabang, Bangkok, Thailand
Dr. P. Chakrabarti, Sir Padampat Singhania University, Udaipur, India
Mr. Nafiz Imtiaz Bin Hamid, Islamic University of Technology (IUT), Bangladesh.
Shahab-A. Shamshirband, Islamic Azad University, Chalous, Iran
Prof. B. Priestly Shan, Anna Univeristy, Tamilnadu, India
Venkatramreddy Velma, Dept. of Bioinformatics, University of Mississippi Medical Center, Jackson MS USA
Akshi Kumar, Dept. of Computer Engineering, Delhi Technological University, India
Dr. Umesh Kumar Singh, Vikram University, Ujjain, India
Mr. Serguei A. Mokhov, Concordia University, Canada
Mr. Lai Khin Wee, Universiti Teknologi Malaysia, Malaysia
Dr. Awadhesh Kumar Sharma, Madan Mohan Malviya Engineering College, India
Mr. Syed R. Rizvi, Analytical Services & Materials, Inc., USA
Dr. S. Karthik, SNS Collegeof Technology, India
Mr. Syed Qasim Bukhari, CIMET (Universidad de Granada), Spain
Mr. A.D.Potgantwar, Pune University, India
Dr. Himanshu Aggarwal, Punjabi University, India
Mr. Rajesh Ramachandran, Naipunya Institute of Management and Information Technology, India
Dr. K.L. Shunmuganathan, R.M.K Engg College , Kavaraipettai ,Chennai
Dr. Prasant Kumar Pattnaik, KIST, India.
Dr. Ch. Aswani Kumar, VIT University, India
Mr. Ijaz Ali Shoukat, King Saud University, Riyadh KSA
Mr. Arun Kumar, Sir Padam Pat Singhania University, Udaipur, Rajasthan
Mr. Muhammad Imran Khan, Universiti Teknologi PETRONAS, Malaysia
Dr. Natarajan Meghanathan, Jackson State University, Jackson, MS, USA
Mr. Mohd Zaki Bin Mas'ud, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia
Prof. Dr. R. Geetharamani, Dept. of Computer Science and Eng., Rajalakshmi Engineering College, India
Dr. Smita Rajpal, Institute of Technology and Management, Gurgaon, India
Dr. S. Abdul Khader Jilani, University of Tabuk, Tabuk, Saudi Arabia
Mr. Syed Jamal Haider Zaidi, Bahria University, Pakistan

Dr. N. Devarajan, Government College of Technology, Coimbatore, Tamilnadu, INDIA
Mr. R. Jagadeesh Kannan, RMK Engineering College, India
Mr. Deo Prakash, Shri Mata Vaishno Devi University, India
Mr. Mohammad Abu Naser, Dept. of EEE, IUT, Gazipur, Bangladesh
Assist. Prof. Prasun Ghosal, Bengal Engineering and Science University, India
Mr. Md. Golam Kaosar, School of Engineering and Science, Victoria University, Melbourne City, Australia
Mr. R. Mohammad Shafi, Madanapalle Institute of Technology & Science, India
Dr. F.Sagayaraj Francis, Pondicherry Engineering College, India
Dr. Ajay Goel, HIET, Kaithal, India
Mr. Nayak Sunil Kashibarao, Bahirji Smarak Mahavidyalaya, India
Mr. Suhas J Manangi, Microsoft India
Dr. Kalyankar N. V., Yeshwant Mahavidyalaya, Nanded, India
Dr. K.D. Verma, S.V. College of Post graduate studies & Research, India
Dr. Amjad Rehman, University Technology Malaysia, Malaysia
Mr. Rachit Garg, L K College, Jalandhar, Punjab
Mr. J. William, M.A.M college of Engineering, Trichy, Tamilnadu, India
Prof. Jue-Sam Chou, Nanhua University, College of Science and Technology, Taiwan
Dr. Thorat S.B., Institute of Technology and Management, India
Mr. Ajay Prasad, Sir Padampat Singhania University, Udaipur, India
Dr. Kamaljit I. Lakhtaria, Atmiya Institute of Technology & Science, India
Mr. Syed Raifiul Hussain, Ahsanullah University of Science and Technology, Bangladesh
Mrs Fazeela Tunnis, Najran University, Kingdom of Saudi Arabia
Mrs Kavita Taneja, Maharishi Markandeshwar University, Haryana, India
Mr. Maniyar Shiraz Ahmed, Najran University, Najran, KSA
Mr. Anand Kumar, AMC Engineering College, Bangalore
Dr. Rakesh Chandra Gangwar, Beant College of Engg. & Tech., Gurdaspur (Punjab) India
Dr. V V Rama Prasad, Sree Vidyanikethan Engineering College, India
Assist. Prof. Neetesh Kumar Gupta, Technocrats Institute of Technology, Bhopal (M.P.), India
Mr. Ashish Seth, Uttar Pradesh Technical University, Lucknow, UP India
Dr. V V S S Balaram, Sreenidhi Institute of Science and Technology, India
Mr Rahul Bhatia, Lingaya's Institute of Management and Technology, India
Prof. Niranjana Reddy. P, KITS, Warangal, India
Prof. Rakesh. Lingappa, Vijetha Institute of Technology, Bangalore, India
Dr. Mohammed Ali Hussain, Nimra College of Engineering & Technology, Vijayawada, A.P., India
Dr. A. Srinivasan, MNM Jain Engineering College, Rajiv Gandhi Salai, Thorapakkam, Chennai
Mr. Rakesh Kumar, M.M. University, Mullana, Ambala, India
Dr. Lena Khaled, Zarqa Private University, Aman, Jordan
Ms. Supriya Kapoor, Patni/Lingaya's Institute of Management and Tech., India
Dr. Tossapon Boongoen, Aberystwyth University, UK
Dr. Bilal Alatas, Firat University, Turkey
Assist. Prof. Jyoti Praaksh Singh, Academy of Technology, India
Dr. Ritu Soni, GNG College, India
Dr. Mahendra Kumar, Sagar Institute of Research & Technology, Bhopal, India.
Dr. Binod Kumar, Lakshmi Narayan College of Tech.(LNCT)Bhopal India
Dr. Muzhir Shaban Al-Ani, Amman Arab University Amman – Jordan
Dr. T.C. Manjunath, ATRIA Institute of Tech, India
Mr. Muhammad Zakarya, COMSATS Institute of Information Technology (CIIT), Pakistan

Assist. Prof. Harmunish Taneja, M. M. University, India
Dr. Chitra Dhawale , SICSR, Model Colony, Pune, India
Mrs Sankari Muthukaruppan, Nehru Institute of Engineering and Technology, Anna University, India
Mr. Aaqif Afzaal Abbasi, National University Of Sciences And Technology, Islamabad
Prof. Ashutosh Kumar Dubey, Trinity Institute of Technology and Research Bhopal, India
Mr. G. Appasami, Dr. Pauls Engineering College, India
Mr. M Yasin, National University of Science and Tech, karachi (NUST), Pakistan
Mr. Yaser Miaji, University Utara Malaysia, Malaysia
Mr. Shah Ahsanul Haque, International Islamic University Chittagong (IIUC), Bangladesh
Prof. (Dr) Syed Abdul Sattar, Royal Institute of Technology & Science, India
Dr. S. Sasikumar, Roever Engineering College
Assist. Prof. Monit Kapoor, Maharishi Markandeshwar University, India
Mr. Nwaocha Vivian O, National Open University of Nigeria
Dr. M. S. Vijaya, GR Govindarajulu School of Applied Computer Technology, India
Assist. Prof. Chakresh Kumar, Manav Rachna International University, India
Mr. Kunal Chadha , R&D Software Engineer, Gemalto, Singapore
Mr. Mueen Uddin, Universiti Teknologi Malaysia, UTM , Malaysia
Dr. Dhuha Basheer abdullah, Mosul university, Iraq
Mr. S. Audithan, Annamalai University, India
Prof. Vijay K Chaudhari, Technocrats Institute of Technology , India
Associate Prof. Mohd Ilyas Khan, Technocrats Institute of Technology , India
Dr. Vu Thanh Nguyen, University of Information Technology, HoChiMinh City, VietNam
Assist. Prof. Anand Sharma, MITS, Lakshmangarh, Sikar, Rajasthan, India
Prof. T V Narayana Rao, HITAM Engineering college, Hyderabad
Mr. Deepak Gour, Sir Padampat Singhania University, India
Assist. Prof. Amutharaj Joyson, Kalasalingam University, India
Mr. Ali Balador, Islamic Azad University, Iran
Mr. Mohit Jain, Maharaja Surajmal Institute of Technology, India
Mr. Dilip Kumar Sharma, GLA Institute of Technology & Management, India
Dr. Debojyoti Mitra, Sir padampat Singhania University, India
Dr. Ali Dehghantanha, Asia-Pacific University College of Technology and Innovation, Malaysia
Mr. Zhao Zhang, City University of Hong Kong, China
Prof. S.P. Setty, A.U. College of Engineering, India
Prof. Patel Rakeshkumar Kantilal, Sankalchand Patel College of Engineering, India
Mr. Biswajit Bhowmik, Bengal College of Engineering & Technology, India
Mr. Manoj Gupta, Apex Institute of Engineering & Technology, India
Assist. Prof. Ajay Sharma, Raj Kumar Goel Institute Of Technology, India
Assist. Prof. Ramveer Singh, Raj Kumar Goel Institute of Technology, India
Dr. Hanan Elazhary, Electronics Research Institute, Egypt
Dr. Hosam I. Faiq, USM, Malaysia
Prof. Dipti D. Patil, MAEER's MIT College of Engg. & Tech, Pune, India
Assist. Prof. Devendra Chack, BCT Kumaon engineering College Dwarahat Almora, India
Prof. Manpreet Singh, M. M. Engg. College, M. M. University, India
Assist. Prof. M. Sadiq ali Khan, University of Karachi, Pakistan
Mr. Prasad S. Halgaonkar, MIT - College of Engineering, Pune, India
Dr. Imran Ghani, Universiti Teknologi Malaysia, Malaysia
Prof. Varun Kumar Kakar, Kumaon Engineering College, Dwarahat, India

Assist. Prof. Nisheeth Joshi, Apaji Institute, Banasthali University, Rajasthan, India
Associate Prof. Kunwar S. Vaisla, VCT Kumaon Engineering College, India
Prof Anupam Choudhary, Bhilai School Of Engg.,Bhilai (C.G.),India
Mr. Divya Prakash Shrivastava, Al Jabal Al garbi University, Zawya, Libya
Associate Prof. Dr. V. Radha, Avinashilingam Deemed university for women, Coimbatore.
Dr. Kasarapu Ramani, JNT University, Anantapur, India
Dr. Anuraag Awasthi, Jayoti Vidyapeeth Womens University, India
Dr. C G Ravichandran, R V S College of Engineering and Technology, India
Dr. Mohamed A. Deriche, King Fahd University of Petroleum and Minerals, Saudi Arabia
Mr. Abbas Karimi, Universiti Putra Malaysia, Malaysia
Mr. Amit Kumar, Jaypee University of Engg. and Tech., India
Dr. Nikolai Stoianov, Defense Institute, Bulgaria
Assist. Prof. S. Ranichandra, KSR College of Arts and Science, Tiruchencode
Mr. T.K.P. Rajagopal, Diamond Horse International Pvt Ltd, India
Dr. Md. Ekramul Hamid, Rajshahi University, Bangladesh
Mr. Hemanta Kumar Kalita , TATA Consultancy Services (TCS), India
Dr. Messaouda Azzouzi, Ziane Achour University of Djelfa, Algeria
Prof. (Dr.) Juan Jose Martinez Castillo, "Gran Mariscal de Ayacucho" University and Acantelys research Group, Venezuela
Dr. Jatinderkumar R. Saini, Narmada College of Computer Application, India
Dr. Babak Bashari Rad, University Technology of Malaysia, Malaysia
Dr. Nighat Mir, Effat University, Saudi Arabia
Prof. (Dr.) G.M.Nasira, Sasurie College of Engineering, India
Mr. Varun Mittal, Gemalto Pte Ltd, Singapore
Assist. Prof. Mrs P. Banumathi, Kathir College Of Engineering, Coimbatore
Assist. Prof. Quan Yuan, University of Wisconsin-Stevens Point, US
Dr. Pranam Paul, Narula Institute of Technology, Agarpara, West Bengal, India
Assist. Prof. J. Ramkumar, V.L.B Janakiammal college of Arts & Science, India
Mr. P. Sivakumar, Anna university, Chennai, India
Mr. Md. Humayun Kabir Biswas, King Khalid University, Kingdom of Saudi Arabia
Mr. Mayank Singh, J.P. Institute of Engg & Technology, Meerut, India
HJ. Kamaruzaman Jusoff, Universiti Putra Malaysia
Mr. Nikhil Patrick Lobo, CADES, India
Dr. Amit Wason, Rayat-Bahra Institute of Engineering & Boi-Technology, India
Dr. Rajesh Shrivastava, Govt. Benazir Science & Commerce College, Bhopal, India
Assist. Prof. Vishal Bharti, DCE, Gurgaon
Mrs. Sunita Bansal, Birla Institute of Technology & Science, India
Dr. R. Sudhakar, Dr.Mahalingam college of Engineering and Technology, India
Dr. Amit Kumar Garg, Shri Mata Vaishno Devi University, Katra(J&K), India
Assist. Prof. Raj Gaurang Tiwari, AZAD Institute of Engineering and Technology, India
Mr. Hamed Taherdoost, Tehran, Iran
Mr. Amin Daneshmand Malayeri, YRC, IAU, Malayer Branch, Iran
Mr. Shantanu Pal, University of Calcutta, India
Dr. Terry H. Walcott, E-Promag Consultancy Group, United Kingdom
Dr. Ezekiel U OKIKE, University of Ibadan, Nigeria
Mr. P. Mahalingam, Caledonian College of Engineering, Oman
Dr. Mahmoud M. A. Abd Ellatif, Mansoura University, Egypt

Prof. Kunwar S. Vaisla, BCT Kumaon Engineering College, India
Prof. Mahesh H. Panchal, Kalol Institute of Technology & Research Centre, India
Mr. Muhammad Asad, Technical University of Munich, Germany
Mr. AliReza Shams Shafiqh, Azad Islamic university, Iran
Prof. S. V. Nagaraj, RMK Engineering College, India
Mr. Ashikali M Hasan, Senior Researcher, CelNet security, India
Dr. Adnan Shahid Khan, University Technology Malaysia, Malaysia
Mr. Prakash Gajanan Burade, Nagpur University/ITM college of engg, Nagpur, India
Dr. Jagdish B.Helonde, Nagpur University/ITM college of engg, Nagpur, India
Professor, Doctor BOUHORMA Mohammed, Univertsity Abdelmalek Essaadi, Morocco
Mr. K. Thirumalaivasan, Pondicherry Engg. College, India
Mr. Umbarkar Anantkumar Janardan, Walchand College of Engineering, India
Mr. Ashish Chaurasia, Gyan Ganga Institute of Technology & Sciences, India
Mr. Sunil Taneja, Kurukshetra University, India
Mr. Fauzi Adi Rafrastra, Dian Nuswantoro University, Indonesia
Dr. Yaduvir Singh, Thapar University, India
Dr. Ioannis V. Koskosas, University of Western Macedonia, Greece
Dr. Vasantha Kalyani David, Avinashilingam University for women, Coimbatore
Dr. Ahmed Mansour Manasrah, Universiti Sains Malaysia, Malaysia
Miss. Nazanin Sadat Kazazi, University Technology Malaysia, Malaysia
Mr. Saeed Rasouli Heikalabad, Islamic Azad University - Tabriz Branch, Iran
Assoc. Prof. Dhirendra Mishra, SVKM's NMIMS University, India
Prof. Shapor Zarei, UAE Inventors Association, UAE
Prof. B.Raja Sarath Kumar, Lenora College of Engineering, India
Dr. Bashir Alam, Jamia millia Islamia, Delhi, India
Prof. Anant J Umbarkar, Walchand College of Engg., India
Assist. Prof. B. Bharathi, Sathyabama University, India
Dr. Fokrul Alom Mazarbhuiya, King Khalid University, Saudi Arabia
Prof. T.S.Jeyali Laseeth, Anna University of Technology, Tirunelveli, India
Dr. M. Balraju, Jawahar Lal Nehru Technological University Hyderabad, India
Dr. Vijayalakshmi M. N., R.V.College of Engineering, Bangalore
Prof. Walid Moudani, Lebanese University, Lebanon
Dr. Saurabh Pal, VBS Purvanchal University, Jaunpur, India
Associate Prof. Suneet Chaudhary, Dehradun Institute of Technology, India
Associate Prof. Dr. Manuj Darbari, BBD University, India
Ms. Prema Selvaraj, K.S.R College of Arts and Science, India
Assist. Prof. Ms.S.Sasikala, KSR College of Arts & Science, India
Mr. Sukhvinder Singh Deora, NC Institute of Computer Sciences, India
Dr. Abhay Bansal, Amity School of Engineering & Technology, India
Ms. Sumita Mishra, Amity School of Engineering and Technology, India
Professor S. Viswanadha Raju, JNT University Hyderabad, India
Mr. Asghar Shahrzad Khashandarag, Islamic Azad University Tabriz Branch, India
Mr. Manoj Sharma, Panipat Institute of Engg. & Technology, India
Mr. Shakeel Ahmed, King Faisal University, Saudi Arabia
Dr. Mohamed Ali Mahjoub, Institute of Engineer of Monastir, Tunisia
Mr. Adri Jovin J.J., SriGuru Institute of Technology, India
Dr. Sukumar Senthilkumar, Universiti Sains Malaysia, Malaysia

Mr. Rakesh Bharati, Dehradun Institute of Technology Dehradun, India
Mr. Shervan Fekri Ershad, Shiraz International University, Iran
Mr. Md. Safiqul Islam, Daffodil International University, Bangladesh
Mr. Mahmudul Hasan, Daffodil International University, Bangladesh
Prof. Mandakini Tayade, UIT, RGTU, Bhopal, India
Ms. Sarla More, UIT, RGTU, Bhopal, India
Mr. Tushar Hrishikesh Jaware, R.C. Patel Institute of Technology, Shirpur, India
Ms. C. Divya, Dr G R Damodaran College of Science, Coimbatore, India
Mr. Fahimuddin Shaik, Annamacharya Institute of Technology & Sciences, India
Dr. M. N. Giri Prasad, JNTUCE,Pulivendula, A.P., India
Assist. Prof. Chintan M Bhatt, Charotar University of Science And Technology, India
Prof. Sahista Machchhar, Marwadi Education Foundation's Group of institutions, India
Assist. Prof. Navnish Goel, S. D. College Of Enginnering & Technology, India
Mr. Khaja Kamaluddin, Sirt University, Sirt, Libya
Mr. Mohammad Zaidul Karim, Daffodil International, Bangladesh
Mr. M. Vijayakumar, KSR College of Engineering, Tiruchengode, India
Mr. S. A. Ahsan Rajon, Khulna University, Bangladesh
Dr. Muhammad Mohsin Nazir, LCW University Lahore, Pakistan
Mr. Mohammad Asadul Hoque, University of Alabama, USA
Mr. P.V.Sarathchand, Indur Institute of Engineering and Technology, India
Mr. Durgesh Samadhiya, Chung Hua University, Taiwan
Dr Venu Kuthadi, University of Johannesburg, Johannesburg, RSA
Dr. (Er) Jasvir Singh, Guru Nanak Dev University, Amritsar, Punjab, India
Mr. Jasmin Cosic, Min. of the Interior of Una-sana canton, B&H, Bosnia and Herzegovina
Dr S. Rajalakshmi, Botho College, South Africa
Dr. Mohamed Sarrab, De Montfort University, UK
Mr. Basappa B. Kodada, Canara Engineering College, India
Assist. Prof. K. Ramana, Annamacharya Institute of Technology and Sciences, India
Dr. Ashu Gupta, Apeejay Institute of Management, Jalandhar, India
Assist. Prof. Shaik Rasool, Shadan College of Engineering & Technology, India
Assist. Prof. K. Suresh, Annamacharya Institute of Tech & Sci. Rajampet, AP, India
Dr . G. Singaravel, K.S.R. College of Engineering, India
Dr B. G. Geetha, K.S.R. College of Engineering, India
Assist. Prof. Kavita Choudhary, ITM University, Gurgaon
Dr. Mehrdad Jalali, Azad University, Mashhad, Iran
Megha Goel, Shamli Institute of Engineering and Technology, Shamli, India
Mr. Chi-Hua Chen, Institute of Information Management, National Chiao-Tung University, Taiwan (R.O.C.)
Assoc. Prof. A. Rajendran, RVS College of Engineering and Technology, India
Assist. Prof. S. Jaganathan, RVS College of Engineering and Technology, India
Assoc. Prof. (Dr.) A S N Chakravarthy, JNTUK University College of Engineering Vizianagaram (State University)
Assist. Prof. Deepshikha Patel, Technocrat Institute of Technology, India
Assist. Prof. Maram Balajee, GMRIT, India
Assist. Prof. Monika Bhatnagar, TIT, India
Prof. Gaurang Panchal, Charotar University of Science & Technology, India
Prof. Anand K. Tripathi, Computer Society of India
Prof. Jyoti Chaudhary, High Performance Computing Research Lab, India
Assist. Prof. Supriya Raheja, ITM University, India

Dr. Pankaj Gupta, Microsoft Corporation, U.S.A.
Assist. Prof. Panchamukesh Chandaka, Hyderabad Institute of Tech. & Management, India
Prof. Mohan H.S, SJB Institute Of Technology, India
Mr. Hossein Malekinezhad, Islamic Azad University, Iran
Mr. Zatin Gupta, Universti Malaysia, Malaysia
Assist. Prof. Amit Chauhan, Phonics Group of Institutions, India
Assist. Prof. Ajal A. J., METS School Of Engineering, India
Mrs. Omowunmi Omobola Adeyemo, University of Ibadan, Nigeria
Dr. Bharat Bhushan Agarwal, I.F.T.M. University, India
Md. Nazrul Islam, University of Western Ontario, Canada
Tushar Kanti, L.N.C.T, Bhopal, India
Er. Aumreesh Kumar Saxena, SIRTs College Bhopal, India
Mr. Mohammad Monirul Islam, Daffodil International University, Bangladesh
Dr. Kashif Nisar, University Utara Malaysia, Malaysia
Dr. Wei Zheng, Rutgers Univ/ A10 Networks, USA
Associate Prof. Rituraj Jain, Vyas Institute of Engg & Tech, Jodhpur – Rajasthan
Assist. Prof. Apoorvi Sood, I.T.M. University, India
Dr. Kayhan Zrar Ghafoor, University Technology Malaysia, Malaysia
Mr. Swapnil Soner, Truba Institute College of Engineering & Technology, Indore, India
Ms. Yogita Gigras, I.T.M. University, India
Associate Prof. Neelima Sadineni, Pydha Engineering College, India Pydha Engineering College
Assist. Prof. K. Deepika Rani, HITAM, Hyderabad
Ms. Shikha Maheshwari, Jaipur Engineering College & Research Centre, India
Prof. Dr V S Giridhar Akula, Avanthi's Scientific Tech. & Research Academy, Hyderabad
Prof. Dr.S.Saravanan, Muthayammal Engineering College, India
Mr. Mehdi Golsorkhtabar Amiri, Islamic Azad University, Iran
Prof. Amit Sadanand Savyanavar, MITCOE, Pune, India
Assist. Prof. P.Oliver Jayaprakash, Anna University,Chennai
Assist. Prof. Ms. Sujata, ITM University, Gurgaon, India
Dr. Asoke Nath, St. Xavier's College, India
Mr. Masoud Rafighi, Islamic Azad University, Iran
Assist. Prof. RamBabu Pemula, NIMRA College of Engineering & Technology, India
Assist. Prof. Ms Rita Chhikara, ITM University, Gurgaon, India
Mr. Sandeep Maan, Government Post Graduate College, India
Prof. Dr. S. Muralidharan, Mepco Schlenk Engineering College, India
Associate Prof. T.V.Sai Krishna, QIS College of Engineering and Technology, India
Mr. R. Balu, Bharathiar University, Coimbatore, India
Assist. Prof. Shekhar. R, Dr.SM College of Engineering, India
Prof. P. Senthilkumar, Vivekanandha Institue of Engineering and Techology for Woman, India
Mr. M. Kamarajan, PSNA College of Engineering & Technology, India
Dr. Angajala Srinivasa Rao, Jawaharlal Nehru Technical University, India
Assist. Prof. C. Venkatesh, A.I.T.S, Rajampet, India
Mr. Afshin Rezakhani Roozbahani, Ayatollah Boroujerdi University, Iran
Mr. Laxmi chand, SCTL, Noida, India
Dr. Dr. Abdul Hannan, Vivekanand College, Aurangabad
Prof. Mahesh Panchal, KITRC, Gujarat
Dr. A. Subramani, K.S.R. College of Engineering, Tiruchengode

Assist. Prof. Prakash M, Rajalakshmi Engineering College, Chennai, India
Assist. Prof. Akhilesh K Sharma, Sir Padampat Singhania University, India
Ms. Varsha Sahni, Guru Nanak Dev Engineering College, Ludhiana, India
Associate Prof. Trilochan Rout, NM Institute of Engineering and Technology, India
Mr. Srikanta Kumar Mohapatra, NMIFT, Orissa, India
Mr. Waqas Haider Bangyal, Iqra University Islamabad, Pakistan
Dr. S. Vijayaragavan, Christ College of Engineering and Technology, Pondicherry, India
Prof. Elboukhari Mohamed, University Mohammed First, Oujda, Morocco
Dr. Muhammad Asif Khan, King Faisal University, Saudi Arabia
Dr. Nagy Ramadan Darwish Omran, Cairo University, Egypt.
Assistant Prof. Anand Nayyar, KCL Institute of Management and Technology, India
Mr. G. Premsankar, Ericsson, India
Assist. Prof. T. Hemalatha, VELS University, India
Prof. Tejaswini Apte, University of Pune, India
Dr. Edmund Ng Giap Weng, Universiti Malaysia Sarawak, Malaysia
Mr. Mahdi Nouri, Iran University of Science and Technology, Iran
Associate Prof. S. Asif Hussain, Annamacharya Institute of technology & Sciences, India
Mrs. Kavita Pabreja, Maharaja Surajmal Institute (an affiliate of GGSIP University), India
Mr. Vorugunti Chandra Sekhar, DA-IICT, India
Mr. Muhammad Najmi Ahmad Zabidi, Universiti Teknologi Malaysia, Malaysia
Dr. Aderemi A. Atayero, Covenant University, Nigeria
Assist. Prof. Osama Sohaib, Balochistan University of Information Technology, Pakistan
Assist. Prof. K. Suresh, Annamacharya Institute of Technology and Sciences, India
Mr. Hassen Mohammed Abdulla Alsaifi, International Islamic University Malaysia (IIUM) Malaysia
Mr. Robail Yasrab, Virtual University of Pakistan, Pakistan
Mr. R. Balu, Bharathiar University, Coimbatore, India
Prof. Anand Nayyar, KCL Institute of Management and Technology, Jalandhar
Assoc. Prof. Vivek S Deshpande, MIT College of Engineering, India
Prof. K. Saravanan, Anna university Coimbatore, India
Dr. Ravendra Singh, MJP Rohilkhand University, Bareilly, India
Mr. V. Mathivanan, IBRA College of Technology, Sultanate of OMAN
Assoc. Prof. S. Asif Hussain, AITS, India
Assist. Prof. C. Venkatesh, AITS, India
Mr. Sami Ulhaq, SZABIST Islamabad, Pakistan
Dr. B. Justus Rabi, Institute of Science & Technology, India
Mr. Anuj Kumar Yadav, Dehradun Institute of technology, India
Mr. Alejandro Mosquera, University of Alicante, Spain
Assist. Prof. Arjun Singh, Sir Padampat Singhania University (SPSU), Udaipur, India
Dr. Smriti Agrawal, JB Institute of Engineering and Technology, Hyderabad
Assist. Prof. Swathi Sambangi, Visakha Institute of Engineering and Technology, India
Ms. Prabhjot Kaur, Guru Gobind Singh Indraprastha University, India
Mrs. Samaher AL-Hothali, Yanbu University College, Saudi Arabia
Prof. Rajneeshkaur Bedi, MIT College of Engineering, Pune, India
Mr. Hassen Mohammed Abdulla Alsaifi, International Islamic University Malaysia (IIUM)
Dr. Wei Zhang, Amazon.com, Seattle, WA, USA
Mr. B. Santhosh Kumar, C S I College of Engineering, Tamil Nadu
Dr. K. Reji Kumar, , N S S College, Pandalam, India

Assoc. Prof. K. Seshadri Sastry, EIILM University, India
Mr. Kai Pan, UNC Charlotte, USA
Mr. Ruikar Sachin, SGGSIET, India
Prof. (Dr.) Vinodani Katiyar, Sri Ramswaroop Memorial University, India
Assoc. Prof., M. Giri, Sreenivasa Institute of Technology and Management Studies, India
Assoc. Prof. Labib Francis Gergis, Misr Academy for Engineering and Technology (MET), Egypt
Assist. Prof. Amanpreet Kaur, ITM University, India
Assist. Prof. Anand Singh Rajawat, Shri Vaishnav Institute of Technology & Science, Indore
Mrs. Hadeel Saleh Haj Aliwi, Universiti Sains Malaysia (USM), Malaysia
Dr. Abhay Bansal, Amity University, India
Dr. Mohammad A. Mezher, Fahad Bin Sultan University, KSA
Assist. Prof. Nidhi Arora, M.C.A. Institute, India
Prof. Dr. P. Suresh, Karpagam College of Engineering, Coimbatore, India
Dr. Kannan Balasubramanian, Mepco Schlenk Engineering College, India
Dr. S. Sankara Gomathi, Panimalar Engineering college, India
Prof. Anil Kumar Suthar, Gujarat Technological University, L.C. Institute of Technology, India
Assist. Prof. R. Hubert Rajan, NOORUL ISLAM UNIVERSITY, India
Assist. Prof. Dr. Jyoti Mahajan, College of Engineering & Technology
Assist. Prof. Homam Reda El-Taj, College of Network Engineering, Saudi Arabia & Malaysia
Mr. Bijan Paul, Shahjalal University of Science & Technology, Bangladesh
Assoc. Prof. Dr. Ch V Phani Krishna, KL University, India
Dr. Vishal Bhatnagar, Ambedkar Institute of Advanced Communication Technologies & Research, India
Dr. Lamri LAOUAMER, Al Qassim University, Dept. Info. Systems & European University of Brittany, Dept. Computer Science, UBO, Brest, France
Prof. Ashish Babanrao Sasankar, G.H.Raisoni Institute Of Information Technology, India
Prof. Pawan Kumar Goel, Shamli Institute of Engineering and Technology, India
Mr. Ram Kumar Singh, S.V Subharti University, India
Assistant Prof. Sunish Kumar O S, Amaljyothi College of Engineering, India
Dr Sanjay Bhargava, Banasthali University, India
Mr. Pankaj S. Kulkarni, AVEW's Shatabdi Institute of Technology, India
Mr. Roohollah Etemadi, Islamic Azad University, Iran
Mr. Oloruntoyin Sefiu Taiwo, Emmanuel Alayande College Of Education, Nigeria
Mr. Sumit Goyal, National Dairy Research Institute, India
Mr Jaswinder Singh Dilawari, Geeta Engineering College, India
Prof. Raghuraj Singh, Harcourt Butler Technological Institute, Kanpur
Dr. S.K. Mahendran, Anna University, Chennai, India
Dr. Amit Wason, Hindustan Institute of Technology & Management, Punjab
Dr. Ashu Gupta, Apeejay Institute of Management, India
Assist. Prof. D. Asir Antony Gnana Singh, M.I.E.T Engineering College, India
Mrs Mina Farmanbar, Eastern Mediterranean University, Famagusta, North Cyprus
Mr. Maram Balajee, GMR Institute of Technology, India
Mr. Moiz S. Ansari, Isra University, Hyderabad, Pakistan
Mr. Adebayo, Olawale Surajudeen, Federal University of Technology Minna, Nigeria
Mr. Jasvir Singh, University College Of Engg., India
Mr. Vivek Tiwari, MANIT, Bhopal, India
Assoc. Prof. R. Navaneethakrishnan, Bharathiyar College of Engineering and Technology, India
Mr. Somdip Dey, St. Xavier's College, Kolkata, India

Mr. Souleymane Balla-Arabé, Xi'an University of Electronic Science and Technology, China
Mr. Mahabub Alam, Rajshahi University of Engineering and Technology, Bangladesh
Mr. Sathyaprakash P., S.K.P Engineering College, India
Dr. N. Karthikeyan, SNS College of Engineering, Anna University, India
Dr. Binod Kumar, JSPM's, Jayawant Technical Campus, Pune, India
Assoc. Prof. Dinesh Goyal, Suresh Gyan Vihar University, India
Mr. Md. Abdul Ahad, K L University, India
Mr. Vikas Bajpai, The LNM IIT, India
Dr. Manish Kumar Anand, Salesforce (R & D Analytics), San Francisco, USA
Assist. Prof. Dheeraj Murari, Kumaon Engineering College, India
Assoc. Prof. Dr. A. Muthukumaravel, VELS University, Chennai
Mr. A. Siles Balasingh, St.Joseph University in Tanzania, Tanzania
Mr. Ravindra Daga Badgujar, R C Patel Institute of Technology, India
Dr. Preeti Khanna, SVKM's NMIMS, School of Business Management, India
Mr. Kumar Dayanand, Cambridge Institute of Technology, India
Dr. Syed Asif Ali, SMI University Karachi, Pakistan
Prof. Pallvi Pandit, Himachal Pradesh University, India
Mr. Ricardo Verschueren, University of Gloucestershire, UK
Assist. Prof. Mamta Juneja, University Institute of Engineering and Technology, Panjab University, India
Assoc. Prof. P. Surendra Varma, NRI Institute of Technology, JNTU Kakinada, India
Assist. Prof. Gaurav Shrivastava, RGPV / SVITS Indore, India
Dr. S. Sumathi, Anna University, India
Assist. Prof. Ankita M. Kapadia, Charotar University of Science and Technology, India
Mr. Deepak Kumar, Indian Institute of Technology (BHU), India
Dr. Dr. Rajan Gupta, GGSIP University, New Delhi, India
Assist. Prof M. Anand Kumar, Karpagam University, Coimbatore, India
Mr. Mr Arshad Mansoor, Pakistan Aeronautical Complex
Mr. Kapil Kumar Gupta, Ansal Institute of Technology and Management, India
Dr. Neeraj Tomer, SINE International Institute of Technology, Jaipur, India
Assist. Prof. Trunal J. Patel, C.G.Patel Institute of Technology, Uka Tarsadia University, Bardoli, Surat
Mr. Sivakumar, Codework solutions, India
Mr. Mohammad Sadegh Mirzaei, PGNR Company, Iran
Dr. Gerard G. Dumancas, Oklahoma Medical Research Foundation, USA
Mr. Varadala Sridhar, Varadhaman College Engineering College, Affiliated To JNTU, Hyderabad
Assist. Prof. Manoj Dhawan, SVITS, Indore
Assoc. Prof. Chitreshh Banerjee, Suresh Gyan Vihar University, Jaipur, India
Dr. S. Santhi, SCSVMV University, India
Mr. Davood Mohammadi Souran, Ministry of Energy of Iran, Iran
Mr. Shamim Ahmed, Bangladesh University of Business and Technology, Bangladesh
Mr. Sandeep Reddivari, Mississippi State University, USA
Assoc. Prof. Ousmane Thiare, Gaston Berger University, Senegal
Dr. Hazra Imran, Athabasca University, Canada
Dr. Setu Kumar Chaturvedi, Technocrats Institute of Technology, Bhopal, India
Mr. Mohd Dilshad Ansari, Jaypee University of Information Technology, India
Ms. Jaspreet Kaur, Distance Education LPU, India
Dr. D. Nagarajan, Salalah College of Technology, Sultanate of Oman
Dr. K.V.N.R.Sai Krishna, S.V.R.M. College, India

Mr. Himanshu Pareek, Center for Development of Advanced Computing (CDAC), India
Mr. Khaldi Amine, Badji Mokhtar University, Algeria
Mr. Mohammad Sadegh Mirzaei, Scientific Applied University, Iran
Assist. Prof. Khyati Chaudhary, Ram-eesh Institute of Engg. & Technology, India
Mr. Sanjay Agal, Pacific College of Engineering Udaipur, India
Mr. Abdul Mateen Ansari, King Khalid University, Saudi Arabia
Dr. H.S. Behera, Veer Surendra Sai University of Technology (VSSUT), India
Dr. Shrikant Tiwari, Shri Shankaracharya Group of Institutions (SSGI), India
Prof. Ganesh B. Regulwar, Shri Shankarprasad Agnihotri College of Engg, India
Prof. Pinnamaneni Bhanu Prasad, Matrix vision GmbH, Germany
Dr. Shrikant Tiwari, Shri Shankaracharya Technical Campus (SSTC), India
Dr. Siddesh G.K., : Dayananada Sagar College of Engineering, Bangalore, India
Dr. Nadir Bouchama, CERIST Research Center, Algeria
Dr. R. Sathishkumar, Sri Venkateswara College of Engineering, India
Assistant Prof (Dr.) Mohamed Moussaoui, Abdelmalek Essaadi University, Morocco
Dr. S. Malathi, Panimalar Engineering College, Chennai, India
Dr. V. Subedha, Panimalar Institute of Technology, Chennai, India
Dr. Prashant Panse, Swami Vivekanand College of Engineering, Indore, India
Dr. Hamza Aldabbas, Al-Balqa'a Applied University, Jordan
Dr. G. Rasitha Banu, Vel's University, Chennai
Dr. V. D. Ambeth Kumar, Panimalar Engineering College, Chennai
Prof. Anuranjan Misra, Bhagwant Institute of Technology, Ghaziabad, India
Ms. U. Sirthuja, PSG college of arts &science, India
Dr. Ehsan Saradar Torshizi, Urmia University, Iran
Dr. Shamneesh Sharma, APG Shimla University, Shimla (H.P.), India
Assistant Prof. A. S. Syed Navaz, Muthayammal College of Arts & Science, India
Assistant Prof. Ranjit Panigrahi, Sikkim Manipal Institute of Technology, Majitar, Sikkim
Dr. Khaled Eskaf, Arab Academy for Science ,Technology & Maritime Transportation, Egypt
Dr. Nishant Gupta, University of Jammu, India
Assistant Prof. Nagarajan Sankaran, Annamalai University, Chidambaram, Tamilnadu, India
Assistant Prof. Tribikram Pradhan, Manipal Institute of Technology, India
Dr. Nasser Lotfi, Eastern Mediterranean University, Northern Cyprus
Dr. R. Manavalan, K S Rangasamy college of Arts and Science, Tamilnadu, India
Assistant Prof. P. Krishna Sankar, K S Rangasamy college of Arts and Science, Tamilnadu, India
Dr. Rahul Malik, Cisco Systems, USA
Dr. S. C. Lingareddy, ALPHA College of Engineering, India
Assistant Prof. Mohammed Shuaib, Interl University, Lucknow, India
Dr. Sachin Yele, Sanghvi Institute of Management & Science, India
Dr. T. Thambidurai, Sun Univercell, Singapore
Prof. Anandkumar Telang, BKIT, India
Assistant Prof. R. Poorvadevi, SCVMV University, India
Dr Uttam Mande, Gitam University, India
Dr. Poornima Girish Naik, Shahu Institute of Business Education and Research (SIBER), India
Prof. Md. Abu Kausar, Jaipur National University, Jaipur, India
Dr. Mohammed Zuber, AISECT University, India
Prof. Kalum Priyanath Udagepola, King Abdulaziz University, Saudi Arabia
Dr. K. R. Ananth, Velalar College of Engineering and Technology, India

Assistant Prof. Sanjay Sharma, Roorkee Engineering & Management Institute Shamli (U.P), India
Assistant Prof. Panem Charan Arur, Priyadarshini Institute of Technology, India
Dr. Ashwak Mahmood muhsen alabaichi, Karbala University / College of Science, Iraq
Dr. Urmila Shrawankar, G H Raisoni College of Engineering, Nagpur (MS), India
Dr. Krishan Kumar Paliwal, Panipat Institute of Engineering & Technology, India
Dr. Mukesh Negi, Tech Mahindra, India
Dr. Anuj Kumar Singh, Amity University Gurgaon, India
Dr. Babar Shah, Gyeongsang National University, South Korea
Assistant Prof. Jayprakash Upadhyay, SRI-TECH Jabalpur, India
Assistant Prof. Varadala Sridhar, Vidya Jyothi Institute of Technology, India
Assistant Prof. Parameshachari B D, KSIT, Bangalore, India
Assistant Prof. Ankit Garg, Amity University, Haryana, India
Assistant Prof. Rajashe Karappa, SDMCET, Karnataka, India
Assistant Prof. Varun Jasuja, GNIT, India
Assistant Prof. Sonal Honale, Abha Gaikwad Patil College of Engineering Nagpur, India
Dr. Pooja Choudhary, CT Group of Institutions, NIT Jalandhar, India
Dr. Faouzi Hidoussi, UHL Batna, Algeria
Dr. Naseer Ali Husieen, Wasit University, Iraq
Assistant Prof. Vinod Kumar Shukla, Amity University, Dubai
Dr. Ahmed Farouk Metwaly, K L University
Mr. Mohammed Noaman Murad, Cihan University, Iraq
Dr. Suxing Liu, Arkansas State University, USA
Dr. M. Gomathi, Velalar College of Engineering and Technology, India
Assistant Prof. Sumardiono, College PGRI Blitar, Indonesia
Dr. Latika Kharb, Jagan Institute of Management Studies (JIMS), Delhi, India
Associate Prof. S. Raja, Pauls College of Engineering and Technology, Tamilnadu, India
Assistant Prof. Seyed Reza Pakize, Shahid Sani High School, Iran
Dr. Thiyyagu Nagaraj, University-INOU, India
Assistant Prof. Noreen Sarai, Harare Institute of Technology, Zimbabwe
Assistant Prof. Gajanand Sharma, Suresh Gyan Vihar University Jaipur, Rajasthan, India
Assistant Prof. Mapari Vikas Prakash, Siddhant COE, Sudumbare, Pune, India
Dr. Devesh Katiyar, Shri Ramswaroop Memorial University, India
Dr. Shenshen Liang, University of California, Santa Cruz, US
Assistant Prof. Mohammad Abu Omar, Limkokwing University of Creative Technology- Malaysia
Mr. Snehasis Banerjee, Tata Consultancy Services, India
Assistant Prof. Kibona Lusekelo, Ruaha Catholic University (RUCU), Tanzania
Assistant Prof. Adib Kabir Chowdhury, University College Technology Sarawak, Malaysia
Dr. Ying Yang, Computer Science Department, Yale University, USA
Dr. Vinay Shukla, Institute Of Technology & Management, India
Dr. Liviu Octavian Mafteiu-Scai, West University of Timisoara, Romania
Assistant Prof. Rana Khudhair Abbas Ahmed, Al-Rafidain University College, Iraq
Assistant Prof. Nitin A. Naik, S.R.T.M. University, India
Dr. Timothy Powers, University of Hertfordshire, UK
Dr. S. Prasath, Bharathiar University, Erode, India
Dr. Ritu Shrivastava, SIRTS Bhopal, India
Prof. Rohit Shrivastava, Mittal Institute of Technology, Bhopal, India
Dr. Gianina Mihai, Dunarea de Jos" University of Galati, Romania

Assistant Prof. Ms. T. Kalai Selvi, Erode Sengunthar Engineering College, India
Assistant Prof. Ms. C. Kavitha, Erode Sengunthar Engineering College, India
Assistant Prof. K. Sinivasamoorthi, Erode Sengunthar Engineering College, India
Assistant Prof. Mallikarjun C Sarsamba Bheemnna Khandre Institute Technology, Bhalki, India
Assistant Prof. Vishwanath Chikaraddi, Veermata Jijabai technological Institute (Central Technological Institute), India
Assistant Prof. Dr. Ikvinderpal Singh, Trai Shatabdi GGS Khalsa College, India
Assistant Prof. Mohammed Noaman Murad, Cihan University, Iraq
Professor Yousef Farhaoui, Moulay Ismail University, Errachidia, Morocco
Dr. Parul Verma, Amity University, India
Professor Yousef Farhaoui, Moulay Ismail University, Errachidia, Morocco
Assistant Prof. Madhavi Dhingra, Amity University, Madhya Pradesh, India
Assistant Prof.. G. Selvavinayagam, SNS College of Technology, Coimbatore, India
Assistant Prof. Madhavi Dhingra, Amity University, MP, India
Professor Kartheesan Log, Anna University, Chennai
Professor Vasudeva Acharya, Shri Madhwa vadira Institute of Technology, India
Dr. Asif Iqbal Hajamydeen, Management & Science University, Malaysia
Assistant Prof., Mahendra Singh Meena, Amity University Haryana
Assistant Professor Manjeet Kaur, Amity University Haryana
Dr. Mohamed Abd El-Basset Matwalli, Zagazig University, Egypt
Dr. Ramani Kannan, Universiti Teknologi PETRONAS, Malaysia
Assistant Prof. S. Jagadeesan Subramaniam, Anna University, India
Assistant Prof. Dharmendra Choudhary, Tripura University, India
Assistant Prof. Deepika Vodnala, SR Engineering College, India
Dr. Kai Cong, Intel Corporation & Computer Science Department, Portland State University, USA
Dr. Kailas R Patil, Vishwakarma Institute of Information Technology (VIIT), India
Dr. Omar A. Alzubi, Faculty of IT / Al-Balqa Applied University, Jordan
Assistant Prof. Kareemullah Shaik, Nimra Institute of Science and Technology, India
Assistant Prof. Chirag Modi, NIT Goa
Dr. R. Ramkumar, Nandha Arts And Science College, India
Dr. Priyadarshini Vydhalingam, Harathiar University, India
Dr. P. S. Jagadeesh Kumar, DBIT, Bangalore, Karnataka
Dr. Vikas Thada, AMITY University, Pachgaon
Dr. T. A. Ashok Kumar, Institute of Management, Christ University, Bangalore
Dr. Shaheera Rashwan, Informatics Research Institute
Dr. S. Preetha Gunasekar, Bharathiyar University, India
Asst Professor Sameer Dev Sharma, Uttarakhand University, Dehradun
Dr. Zhihan Lv, Chinese Academy of Science, China
Dr. Ikvinderpal Singh, Trai Shatabdi GGS Khalsa College, Amritsar
Dr. Umar Ruhi, University of Ottawa, Canada
Dr. Jasmin Cosic, University of BiHac, Bosnia and Herzegovina
Dr. Homam Reda El-Taj, University of Tabuk, Kingdom of Saudi Arabia
Dr. Mostafa Ghobaei Arani, Islamic Azad University, Iran
Dr. Ayyasamy Ayyanar, Annamalai University, India
Dr. Selvakumar Manickam, Universiti Sains Malaysia, Malaysia
Dr. Murali Krishna Namana, GITAM University, India
Dr. Smriti Agrawal, Chaitanya Bharathi Institute of Technology, Hyderabad, India
Professor Vimalathithan Rathinasabapathy, Karpagam College Of Engineering, India

Dr. Sushil Chandra Dimri, Graphic Era University, India
Dr. Dinh-Sinh Mai, Le Quy Don Technical University, Vietnam
Dr. S. Rama Sree, Aditya Engg. College, India
Dr. Ehab T. Alnfrawy, Sadat Academy, Egypt
Dr. Patrick D. Cerna, Haramaya University, Ethiopia
Dr. Vishal Jain, Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), India
Associate Prof. Dr. Jiliang Zhang, North Eastern University, China
Dr. Sharefa Murad, Middle East University, Jordan
Dr. Ajeet Singh Poonia, Govt. College of Engineering & technology, Rajasthan, India
Dr. Vahid Esmaeelzadeh, University of Science and Technology, Iran
Dr. Jacek M. Czerniak, Casimir the Great University in Bydgoszcz, Institute of Technology, Poland
Associate Prof. Anisur Rehman Nasir, Jamia Millia Islamia University
Assistant Prof. Imran Ahmad, COMSATS Institute of Information Technology, Pakistan
Professor Ghulam Qasim, Preston University, Islamabad, Pakistan
Dr. Parameshachari B D, GSSS Institute of Engineering and Technology for Women
Dr. Wencan Luo, University of Pittsburgh, US
Dr. Musa PEKER, Faculty of Technology, Mugla Sitki Kocman University, Turkey
Dr. Gunasekaran Shanmugam, Anna University, India
Dr. Binh P. Nguyen, National University of Singapore, Singapore
Dr. Rajkumar Jain, Indian Institute of Technology Indore, India
Dr. Imtiaz Ali Halepoto, QUEST Nawabshah, Pakistan
Dr. Shaligram Prajapat, Devi Ahilya University Indore India
Dr. Sunita Singhal, Birla Institute of Technologyand Science, Pilani, India
Dr. Ijaz Ali Shoukat, King Saud University, Saudi Arabia
Dr. Anuj Gupta, IKG Punjab Technical University, India
Dr. Sonali Saini, IES-IPS Academy, India
Dr. Krishan Kumar, MotiLal Nehru National Institute of Technology, Allahabad, India
Dr. Z. Faizal Khan, College of Engineering, Shaqra University, Kingdom of Saudi Arabia
Prof. M. Padmavathamma, S.V. University Tirupati, India
Prof. A. Velayudham, Cape Institute of Technology, India
Prof. Seifeide Kadry, American University of the Middle East
Dr. J. Durga Prasad Rao, Pt. Ravishankar Shukla University, Raipur
Assistant Prof. Najam Hasan, Dhofar University
Dr. G. Suseendran, Vels University, Pallavaram, Chennai
Prof. Ankit Faldu, Gujarat Technological Universiry- Atmiya Institute of Technology and Science
Dr. Ali Habiboghli, Islamic Azad University
Dr. Deepak Dembla, JECRC University, Jaipur, India
Dr. Pankaj Rajan, Walmart Labs, USA
Assistant Prof. Radoslava Kraleva, South-West University "Neofit Rilski", Bulgaria
Assistant Prof. Medhavi Shriwas, Shri vaishnav institute of Technology, India
Associate Prof. Sedat Akleylek, Ondokuz Mayis University, Turkey
Dr. U.V. Arivazhagu, Kingston Engineering College Affiliated To Anna University, India
Dr. Touseef Ali, University of Engineering and Technology, Taxila, Pakistan
Assistant Prof. Naren Jeeva, SASTRA University, India
Dr. Riccardo Colella, University of Salento, Italy
Dr. Enache Maria Cristina, University of Galati, Romania
Dr. Senthil P, Kurinji College of Arts & Science, India

- Dr. Hasan Ashrafi-rizi, Isfahan University of Medical Sciences, Isfahan, Iran
Dr. Mazhar Malik, Institute of Southern Punjab, Pakistan
Dr. Yajie Miao, Carnegie Mellon University, USA
Dr. Kamran Shaukat, University of the Punjab, Pakistan
Dr. Sasikaladevi N., SASTRA University, India
Dr. Ali Asghar Rahmani Hosseiniabadi, Islamic Azad University Ayatollah Amoli Branch, Amol, Iran
Dr. Velin Kralev, South-West University "Neofit Rilski", Blagoevgrad, Bulgaria
Dr. Marius Iulian Mihailescu, LUMINA - The University of South-East Europe
Dr. Sriramula Nagaprasad, S.R.R.Govt.Arts & Science College, Karimnagar, India
Prof (Dr.) Namrata Dhanda, Dr. APJ Abdul Kalam Technical University, Lucknow, India
Dr. Javed Ahmed Maher, Shah Abdul Latif University, Khairpur Mir's, Pakistan
Dr. B. Narendra Kumar Rao, Sree Vidyanikethan Engineering College, India
Dr. Shahzad Anwar, University of Engineering & Technology Peshawar, Pakistan
Dr. Basit Shahzad, King Saud University, Riyadh - Saudi Arabia
Dr. Nilamadhab Mishra, Chang Gung University
Dr. Sachin Kumar, Indian Institute of Technology Roorkee
Dr. Santosh Nanda, Biju-Pattnaik University of Technology
Dr. Sherzod Turaev, International Islamic University Malaysia
Dr. Yilun Shang, Tongji University, Department of Mathematics, Shanghai, China
Dr. Nuzhat Shaikh, Modern Education society's College of Engineering, Pune, India
Dr. Parul Verma, Amity University, Lucknow campus, India
Dr. Rachid Alaoui, Agadir Ibn Zohr University, Agadir, Morocco
Dr. Dharmendra Patel, Charotar University of Science and Technology, India
Dr. Dong Zhang, University of Central Florida, USA
Dr. Kennedy Chinedu Okafor, Federal University of Technology Owerri, Nigeria
Prof. C Ram Kumar, Dr NGP Institute of Technology, India
Dr. Sandeep Gupta, GGS IP University, New Delhi, India
Dr. Shahanawaj Ahamad, University of Ha'il, Ha'il City, Ministry of Higher Education, Kingdom of Saudi Arabia
Dr. Najeeb Ahmed Khan, NED University of Engineering & Technology, India
Dr. Sajid Ullah Khan, Universiti Malaysia Sarawak, Malaysia
Dr. Muhammad Asif, National Textile University Faisalabad, Pakistan
Dr. Yu BI, University of Central Florida, Orlando, FL, USA
Dr. Brijendra Kumar Joshi, Research Center, Military College of Telecommunication Engineering, India
Prof. Dr. Nak Eun Cho, Pukyong National University, Korea
Prof. Wasim Ul-Haq, Mathematics Department Faculty of Science, Majmaah University, Saudi Arabia

CALL FOR PAPERS

International Journal of Computer Science and Information Security

IJCSIS 2017-2018

ISSN: 1947-5500

<http://sites.google.com/site/ijcsis/>

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc.,) Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on

its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embedded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at <http://sites.google.com/site/ijcsis/authors-notes>.

**© IJCSIS PUBLICATION 2017
ISSN 1947 5500
<http://sites.google.com/site/ijcsis/>**