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

International Journal of Computer Science & Information Security

Pennsylvania, USA

Indexed and technically co-sponsored by :

IJCSIS

ISSN (online): 1947-5500

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 Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.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/

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):

Bibliographic Information
ISSN: 1947-5500
Monthly publication (Regular Special Issues)
Commenced Publication since May 2009

Editorial / Paper Submissions:
IJCSIS Managing Editor
(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 Enigneering, 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 Esmaeelzadeh [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 Akleylek [Profile] Ondokuz Mayis University, Turkey	**Dr. Sachin Kumar [Profile]** Indian Institute of Technology (IIT) Roorkee

ISSN 1947 5500

TABLE OF CONTENTS

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 UniversitySuwon, 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]

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]

A comprehensive approach on RLE and ECC (Elliptical Cure 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.

$$\textbf{Compression ratio} = \frac{\text{Length of compressed data}}{\text{Length of original data}}$$

$$\textbf{MSE} = \frac{\text{Sum (Data – recover message)}^2}{\text{Count}}$$

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

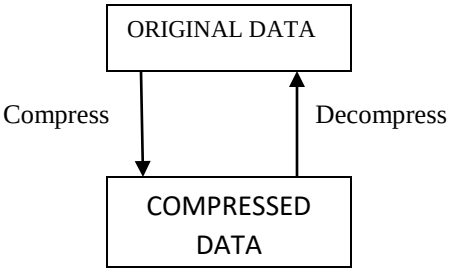


Fig.1 Working of RLE

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.

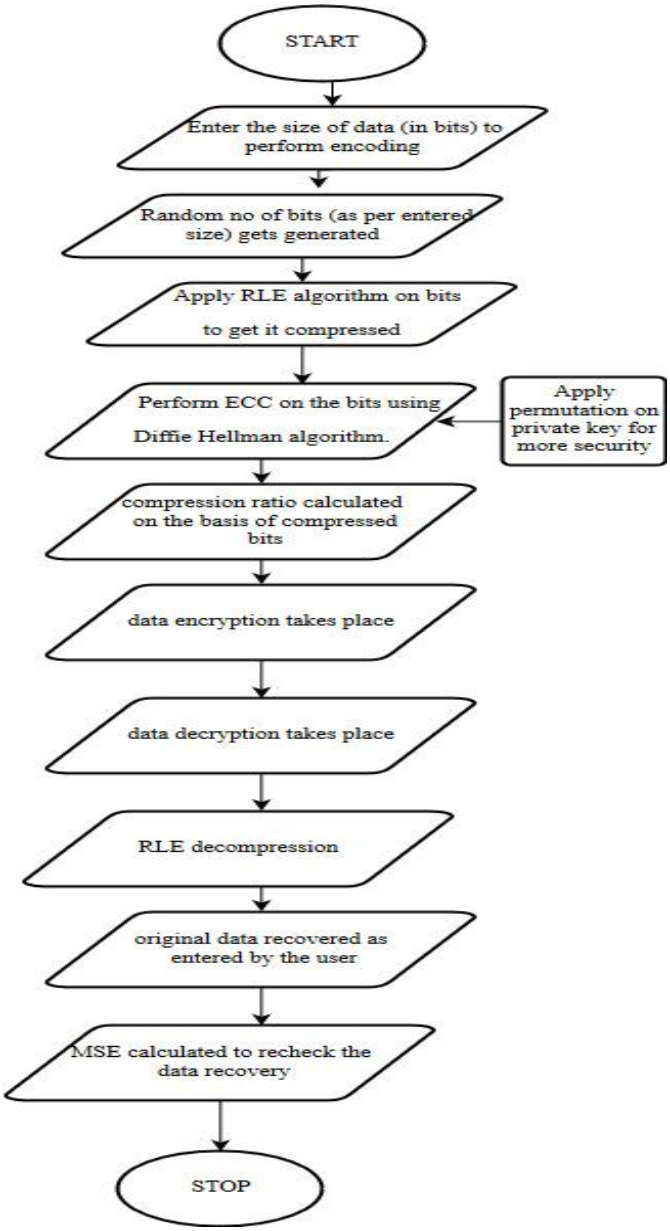


Fig.2. Complete flow of proposed system

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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)).

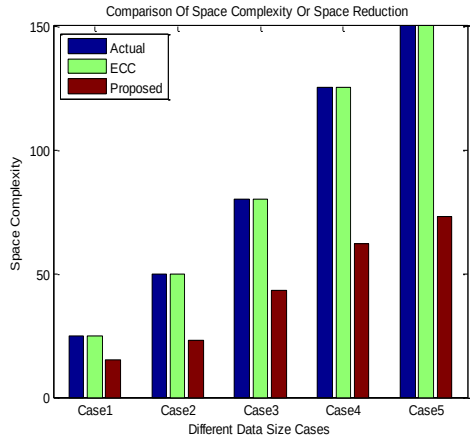


Fig.3. Data compression ratio graph

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.

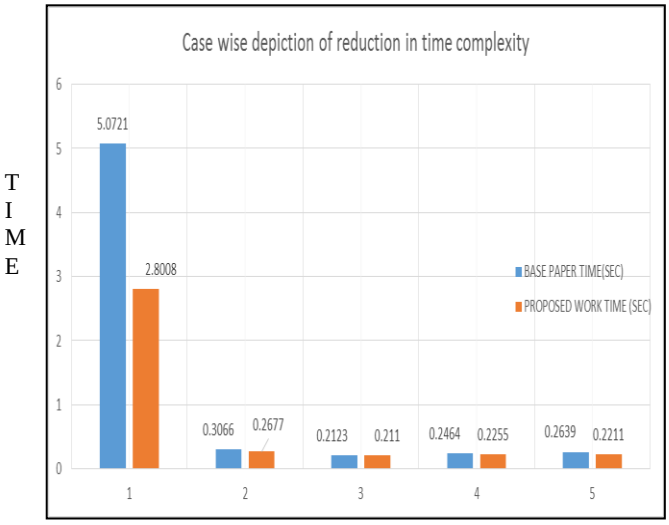


Fig.4 Case wise depiction of reduction in time complexity

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

[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 (ICCSP)*, Melmaruvathur, 2015, pp. 0746-0749.doi: 10.1109/ICCSP.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 Nimbhorkar, 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 Jyotiyana, 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 Jyotiyana,Varun Prakash Saxena (Dec 23-25 2016);Fault attack for scalar multiplication over finite field (E(Fq)) on

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.00,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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

(ICACDOT), Pune, 2016, pp. 13-18. doi:10.1109/ICACDOT.2016.7877543URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7877543&isnumber=787754 0

[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.2017.7950418URL: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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

ENHANCED MULTI-MODAL BIOMETRIC BASED SECURITY SCHEME WITH FEATURE BASED MACHINE LEARNING APPROACH

Rakhi Choudhary[1], Er. Krishan Kumar[2],Dr. Himanshu Monga[3]

M.Tech (Scholar), Assistant Professor, Principal

Department of Computer Science Engineering

Jan Nayak Chaudhary Devi Lal Memorial College of Engineering

Rakhichoudhary930@gmail.com[1], chhaperwal@gmail.com[2], himanshumonga@gmail.com[3]

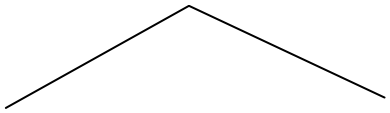
Abstract **–**The multi-modal biometric system, which defined that the behavioural and psychological features to verify a user. In real-time applications mostly used a uni-modal biometric system to security or authenticate user. An identification based on various biometrics represents a developing feature. 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 form 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 conjuction 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].

Biometric Modalities



Physiological **Behavioral**

Fingerprint	**Signature**
Veins	**Voice**
Hand	**Gait**
Eye	**Keystroke**

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 Y is calculated and the sum is subtracted from the considered data set to create it zero sum.

$$\boldsymbol{Yc = Y - F\{Y\}}$$

2. ***Cleaning the data :*** Covariance Matrix Cov mat Y is center of the data Yc is calculated. The E decomposed of the covY is evaluated. If DD is the E matrix and Eigen Values is the V vector of the eigen matrix then

$$\boldsymbol{X = DD - \frac{1}{2} E(Yc)}$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

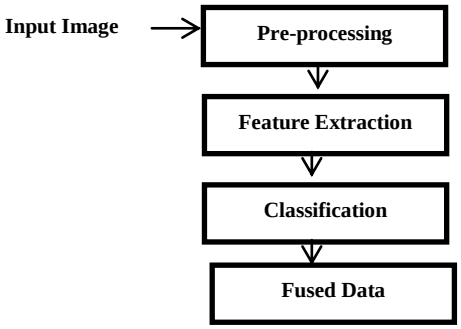


Fig 1. Proposed Model in Multimodel Biometric process

3. ***Fix point repetation 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.

(i) (ii)

(iii)

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.

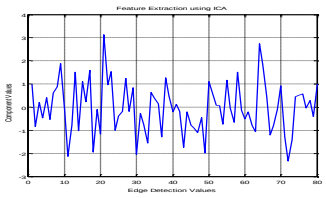


Fig 3. Feature Extraction with ICA

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.

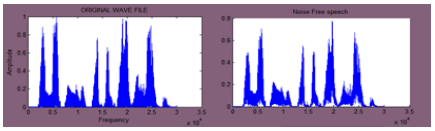
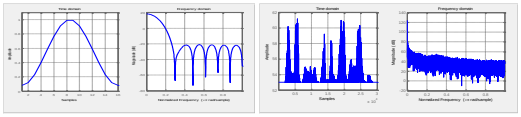
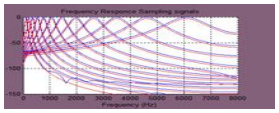


Fig 4. Sample Speech and Noise Free Signal

The above figure shows that the speech recognition module , first upload the wave form, feature extract using GTCC algorithm and fused the multimodle biometric with score level fusion method save in the data matrix form.



(i) (ii)



(iii)

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.

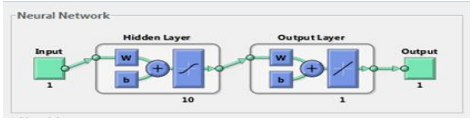


Fig 6. Convolutional Neural Network

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

validate the system to identify how accurately work through performance parameters.

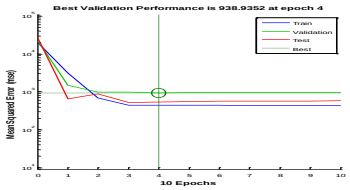


Fig 7. Best Performance

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.

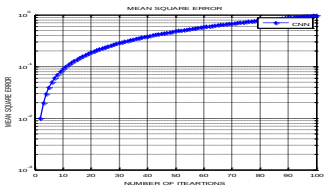
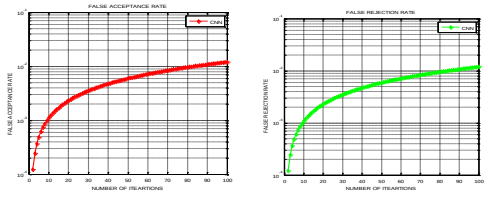


Fig 8. Means Square Error Rate

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.

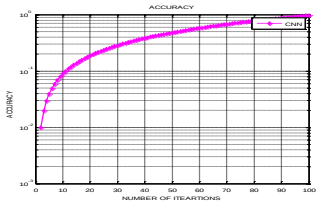


Fig 11. Accuracy in CNN

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%.

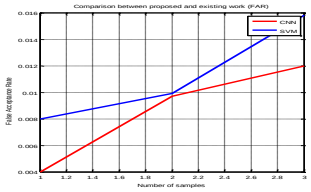


Fig 12. Comparison between FAR with proposed and existing work

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

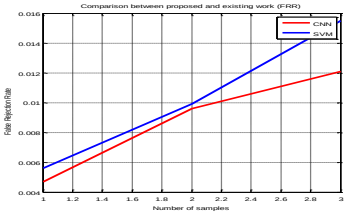


Fig 13. Comparison between FRR proposed and Existing work

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 intelligence28, 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 indescernibility 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

$$N_{R1} = \sum_n \sum \frac{S_1}{D_1} \quad \text{for image 1} \quad (3)$$

$$\text{And } N_{R2} = \sum_n \sum \frac{S_1}{D_2} \quad \text{for image 2} \quad (4)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

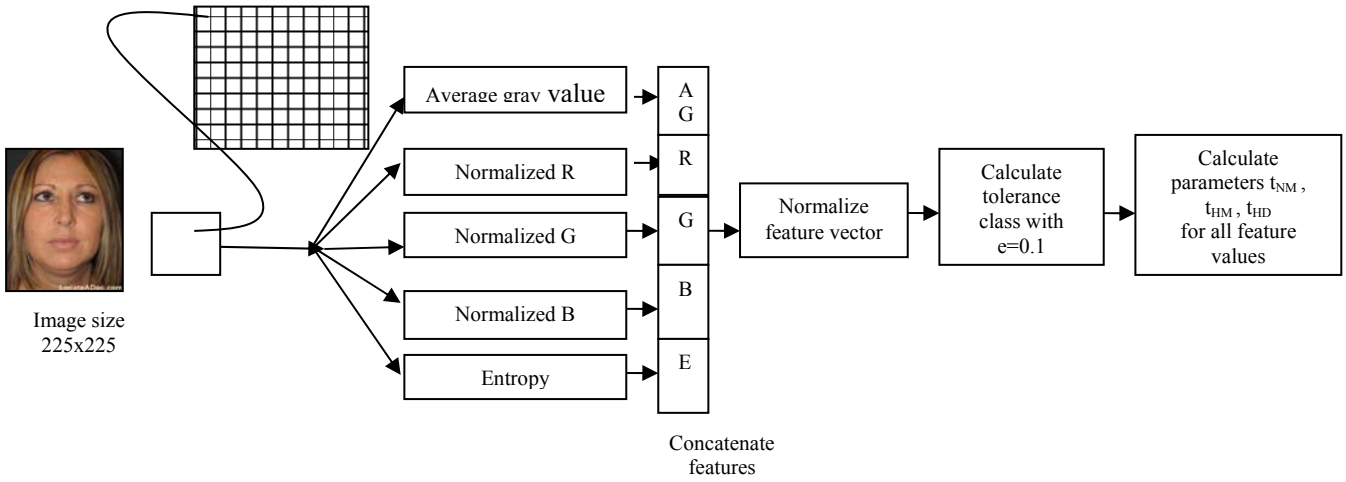


Figure 1 Block diagram of the proposed method using near set theory

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 $1\text{-}e_1$ rather than $1/e_1$ In this case the mathematical new definition of entropy is

$$E = \sum \left[e_1 * e\left(1 - {}^{e_1}/_{255}\right) \right] \quad (9)$$

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 $n_X > n_Y$

Tc ratio = O_{inY} / O_{inX}

Else

Tc ratio = O_{inX}/O_{inY}

Were S_x=number of features of image=81

Objects=25x25=225

$$n_x = length\left(find(c \leq S_x)\right); \quad (11)$$

Were c is row vector of tolerance matrix. Similar way n_Y is calculated for image 2 with

$$n_y = length\left(find(c > S_x)\right); \quad (12)$$

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 T_c 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,

$$t_{NM} = sum[T_1(:,1) * T_1(:,4)')]/sum[T_1(:,1)] \quad (13)$$

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 $t = sum(T_x)$; 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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=XUY) 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.

Finally $t_{HD} = 1 - HD(X,Y)$ (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

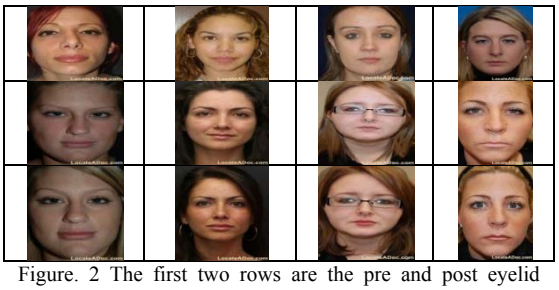


Figure. 2 The first two rows are the pre and post eyelid surgery images(Blepharoplasty) , third and fourth row indicates pre and post nose surgery images(Rhinoplasty) and fifth and sixth row indicates pre and post lips surgery images

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Fig.3 Target before surgery image (Blepharoplasty)

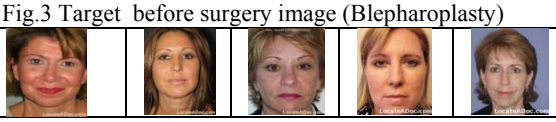


Figure. 4 Sample of after surgery images to be compared (Blepharoplasty)

Figure.5 Sample of Before and after lips surgery images.

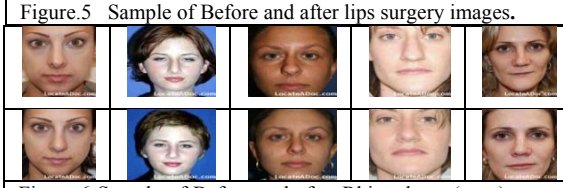


Figure.6 Sample of Before and after Rhinoplasty (nose) surgery images.

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. Romdhami 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. Chellapa, "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. Chellapa, "Facial similarity across age, disguise, illumination and pose",

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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, Yiying. 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. Nooreyezdan, "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", Proceeedings 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,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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):

$$ii(x, y) = \sum i(x', y') \quad \ldots\ldots (1)$$

$$\mathbf{x' \leq x,\ y' \leq y}$$

Where ii (x, y) is the integral image and i (x, y) is the original image.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- 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].

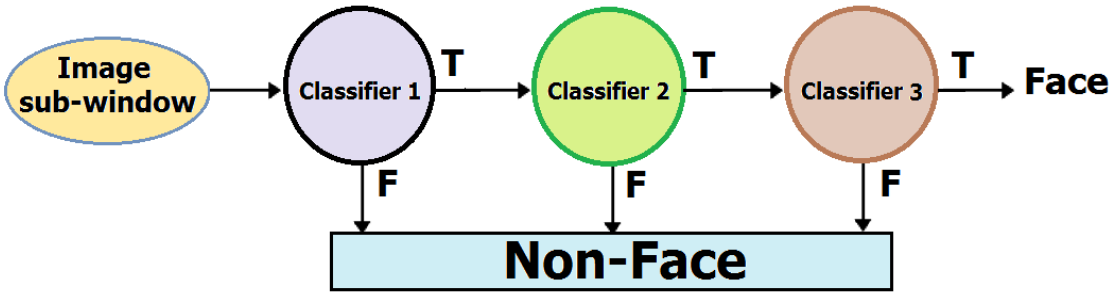


Figure (1) Cascade of Stages

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 ($\mathbf{S}_B$) and the sum of their projected inside class scatter matrix ($\mathbf{S}_W$) as stated by [10][11].

$$J(W) = \frac{|S_B|}{|S_W|} = \frac{|W^T S_B W|}{|W^T S_W W|} = [w_1, \ldots, w_k] \qquad \ldots\ldots\ldots\ldots(2)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Where [w_1,…, w_k] is the eigenvectors

$$S_B = \sum_{j=1}^{c} n_j \ (m_j - m)(m_j - m)^T , \qquad \ldots\ldots\ldots\ldots (3)$$

$$S_W = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{ji} - m_j)(x_{ji} - m_j)^T \ , \qquad \ldots\ldots\ldots\ldots (4)$$

Where c is the number of classes is number of samples in the class, is the ith sample in the class j, *m* is the mean of all classes, *mj* 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].

$$D = \sqrt{\sum_{i=1}^{M} (y_i^A - y_i^B \ )^2} \qquad \ldots\ldots\ldots\ldots (5)$$

And by calculating the $(c - 1)$ Eigen vectors (**w**) and Eigen values (λ) of ($\mathbf{S}^{-1}$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 pa ∈ [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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

For simplicity, this last assumption can be approximated by a fraction pa 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

$$x^{(t+1)}\, i = x^{(t)} i + \alpha \oplus L´evy(\lambda), \qquad \ldots\ldots\ldots\ldots\ldots (6)$$

where _ > 0 is the step size which should be related to the scales of the problem of interest. In most cases, we can use α = O(1). The product _ 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

$$L´evy \_ u = t-\_, (1 < \lambda \leq 3), \quad \ldots\ldots\ldots\ldots (7)$$

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 (x1, y1), (x2, y2),:::: (X | D |, y | D |), where xi is a set of exercises with associated class descriptions, yi. Each yi 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, A1 and A2, 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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:

$$W. X + b = 0 \qquad \text{…….. (8)}$$

Where W is the weight vector, which is W = {w_1, w_2, … 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

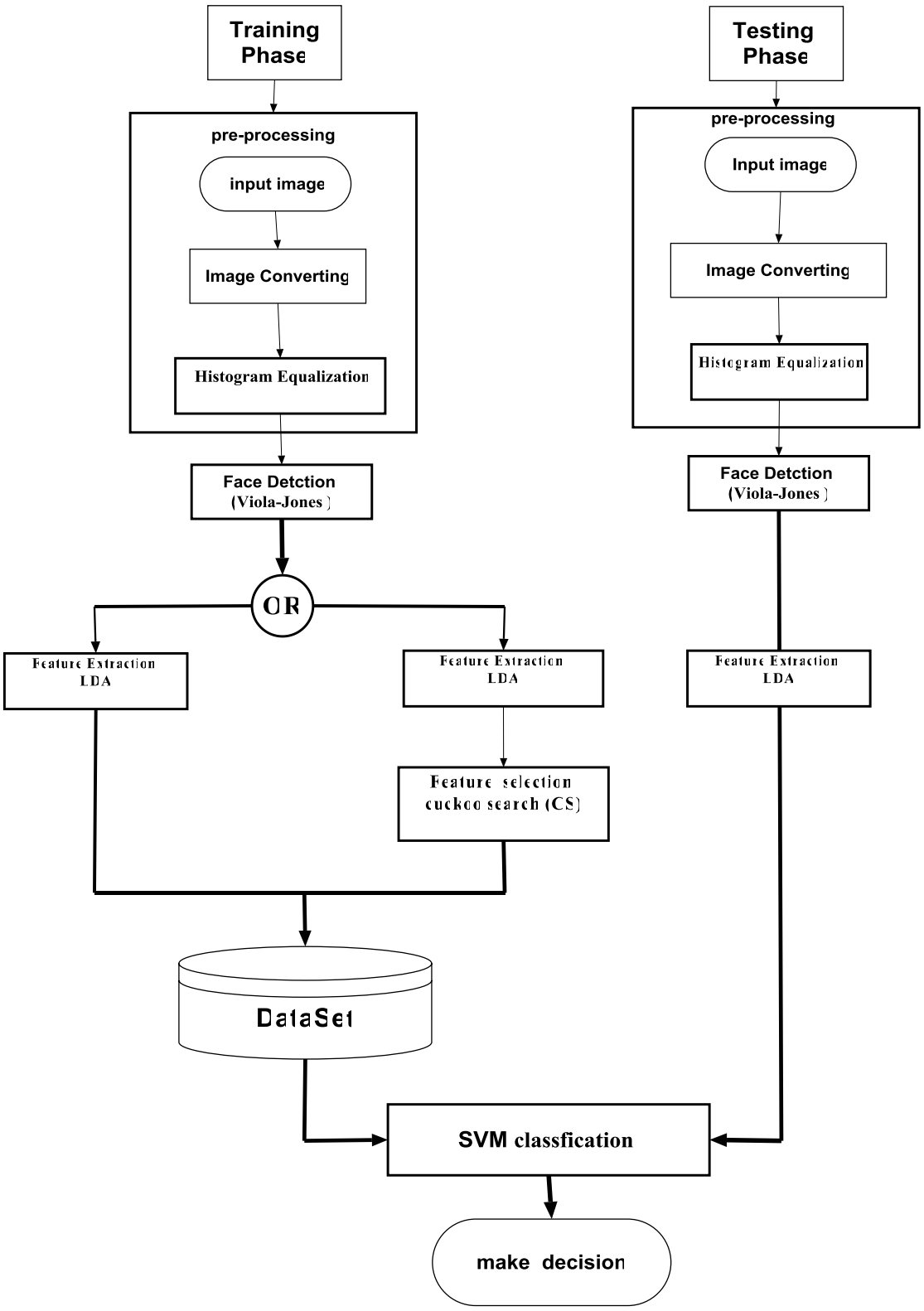


Figure (2): Proposed Face Recognition System

- 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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:

Grayscale value = 0.2125R +0.0715G+ 0.722B …..…(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:

$$S_k = T(r_k) = \sum_{j=0}^{k} n1 / n \qquad \ldots\ldots (9)$$

Whereas k=0, 1, 2… 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 'rk' 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.

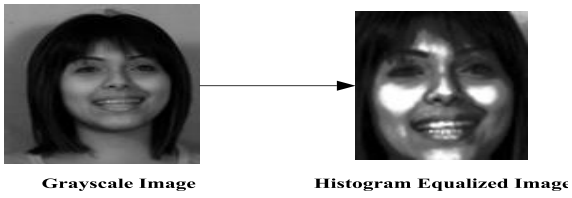
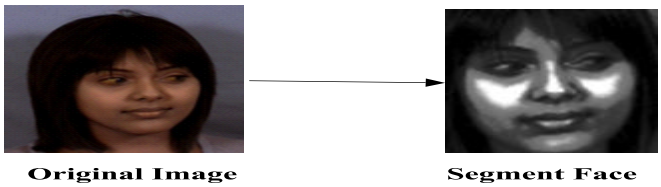


Figure (5): Histogram Equalization

First Phase: segment dataset into four segment consist of face, eye , noise and mouth

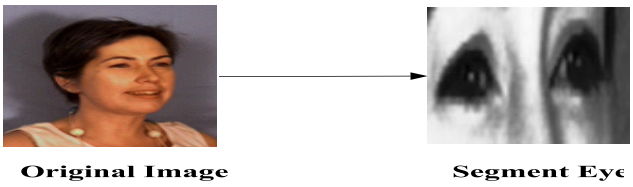
- Face segment image

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



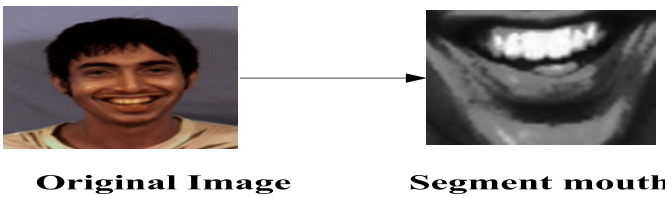
Figure(6): Face Segment Image

- 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 algorithm in face recognition , in thesis getting on 19 featers 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 as following algorithm :

Algorithm1: Cuckoo Search(CS)

Objective function f(x), x = (x1, ..., x_d) 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 > F),
Replace j by the new solution;
End
Abandon a fraction (p_a) 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 applied feature extraction by using LDA then using feature selection by cuckoo search to selection best feature of each image. In testing phase, the face classifier model used to inspect a new face image which does not available in the training set. Classifier

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

$$=\frac{Total\ Number\ of\ sample - l\ Number\ of\ sample\ that\ False\ accepte}{Total\ Number\ of\ Sample} * 100\%$$

LDA	Dataset	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% .

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

cuckoo	Dataset	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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Manufacturing Technology (2013): 1-7.

16. Gandomi, Amir Hossein, et al. "Cuckoo search algorithm: a metaheuristic pproach 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 Appling 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;

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

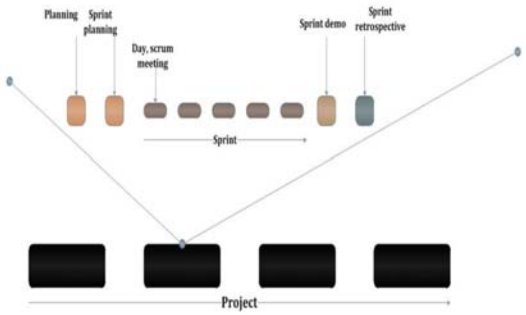


Fig (1): General Structure of the scrum process.

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**

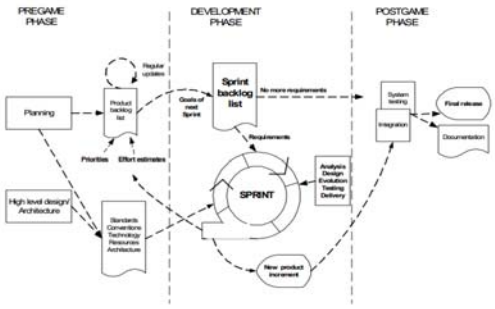


Fig (2): Scrum Process

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).

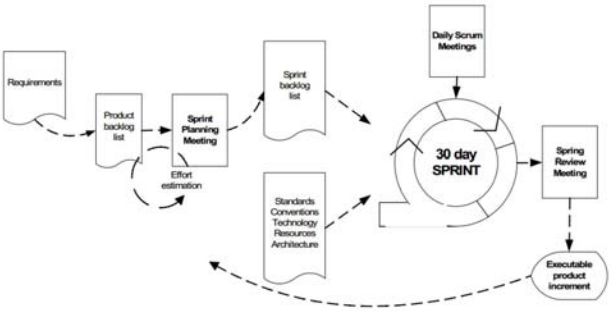


Fig (3): Sprint cycle

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).

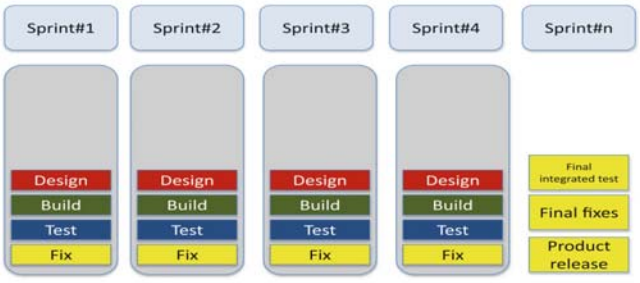


Fig (4): Sprint leading to final product release.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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).

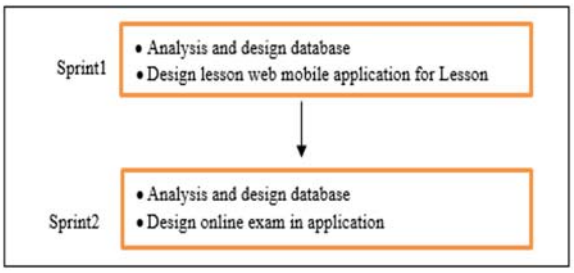


Fig (5): Sprints for system Covering Development from Initiation to a Finished.

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.

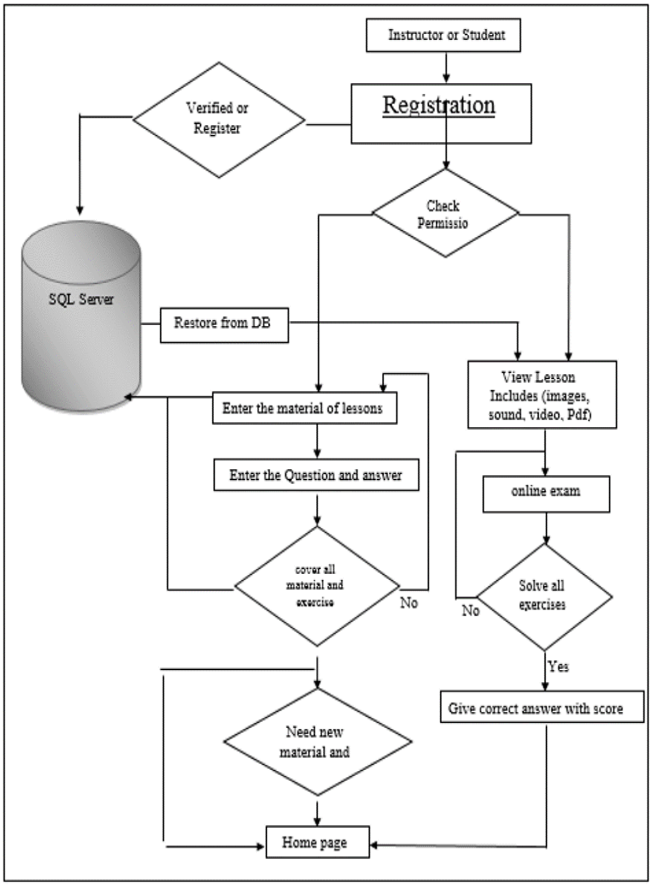


Fig (6): Architecture of proposed system

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table (1): product backlog for proposed system**.**

NO	User story	iteration	Task	complexity	Estimation (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 element so that lesson 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 exam online so that student can take test	2	1-design tables for test, subject and question.	Large	5	9	45	5/9/2016	10/9/2016
11		2	2- built interfaces for test, subject and question in website	Large	5	9	45	11/9/2016	15/9/2016
12	as instructor, I want view result and 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.

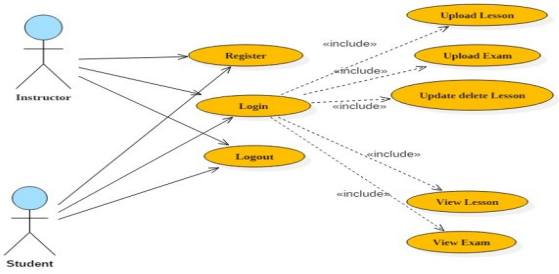


Fig (7): Use case Diagram for proposed system

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 Estamation in Agile Methodology

The differerance between the estamation 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]:-

$$\textit{Relative error: } \mathrm{RE} = \frac{\text{Actual} - \text{Estimate}}{\text{Actual}}$$

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table (2): Estmation 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	

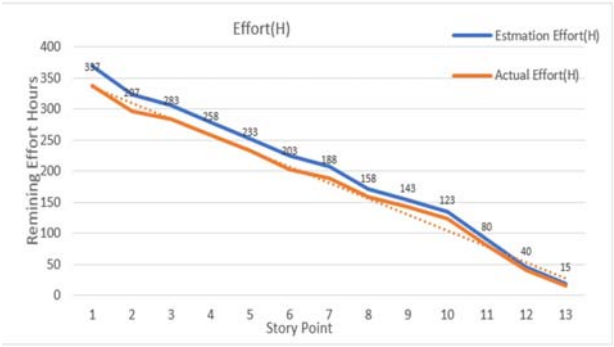


Fig (10): Estmation and actual hours for system.

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

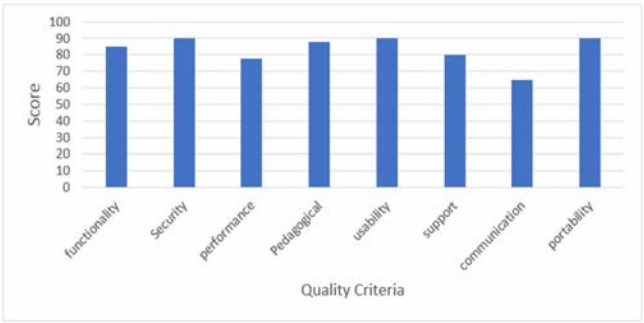


Fig (11): Overall Score for each Quilty Criteria in sytem.

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.

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)

More than 30 papers have been published in the area of interest

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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
[1] maasiddiqui@kau.edu.sa
[2] ssalehalkatheri@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 ennountered. 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 byte sequences and PE header	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 byte sequences	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, Baggin	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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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, CFG, disassembled OpCodes, DIT, miscellaneous files information, and SCT	780 malware and 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 classi?er	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

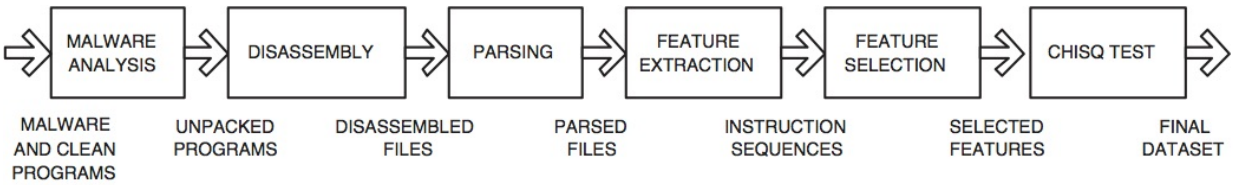


Fig. 1: Data preprocessing steps.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE II: Packers/Compilers Analysis Details of Worms and Clean Programs

	Before Unpacking		After Unpacking	
Packer/Compiler	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

	Before Unpacking		After Unpacking	
Packer/Compiler	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      short near ptr loc_16E+1
dec     sp
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 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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, ... p_n, and m instruction sequences s_1, s_2, ..., 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, ... 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 showed 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 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*.

TABLE IV: Features Statistics

	No of Features			
Feature Size	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

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].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

$DetectionRate = \frac{TP}{TP+FN}$

False Alarm Rate: Percentage of wrongly identified benign programs.

$FalseAlarmRate = \frac{FP}{TN+FP}$

Overall Accuracy: Percentage of correctly identified programs.

$OverallAccuracy = \frac{TP+TN}{TP+TN+FP+FN}$

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

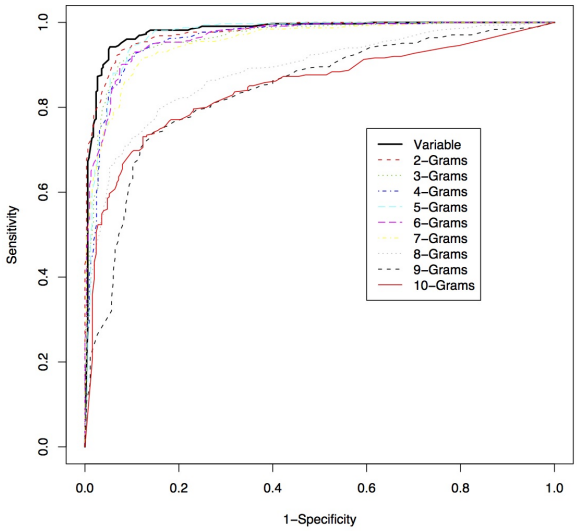


Fig. 4: ROC curves comparing random forest results for each n-gram size using all variables.

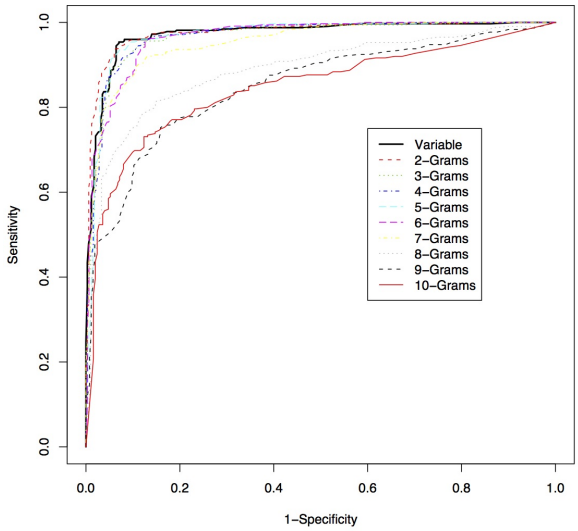


Fig. 5: ROC curves comparing random forest results for each n-gram size using random forest variable selection.

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.

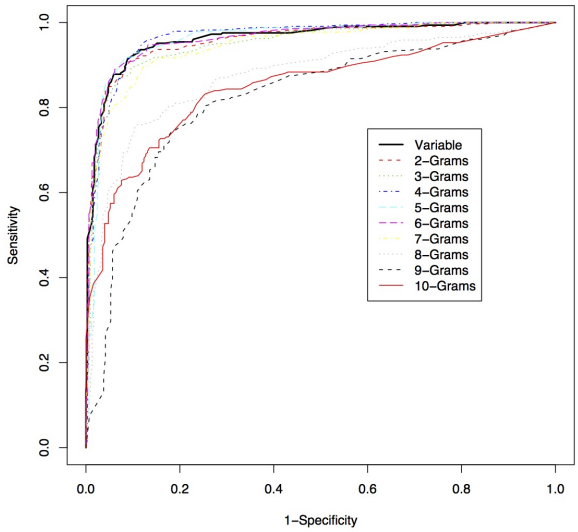


Fig. 6: ROC curves comparing random forest results for each n-gram size using PCA variable selection.

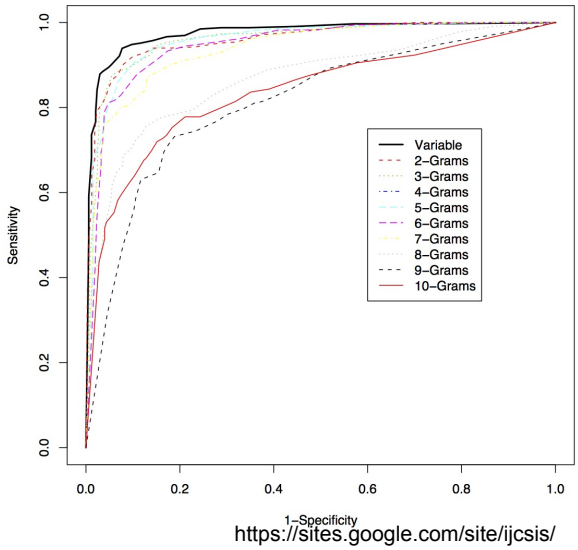


Fig. 7: ROC curves comparing bagging results for each n-gram size using all variables.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

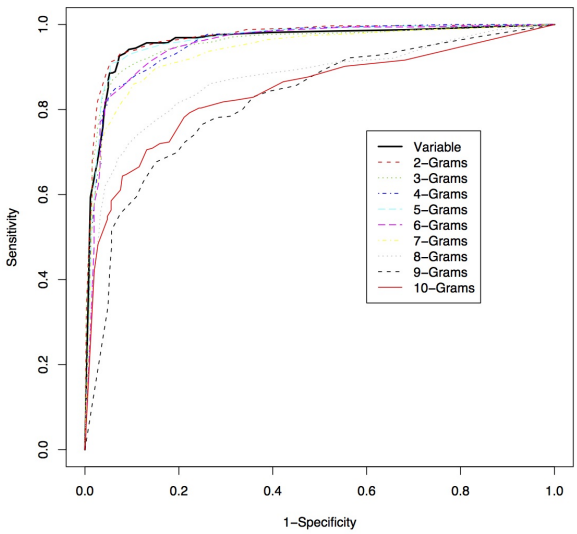


Fig. 8: ROC curves comparing bagging results for each n-gram size using random forest variable selection.

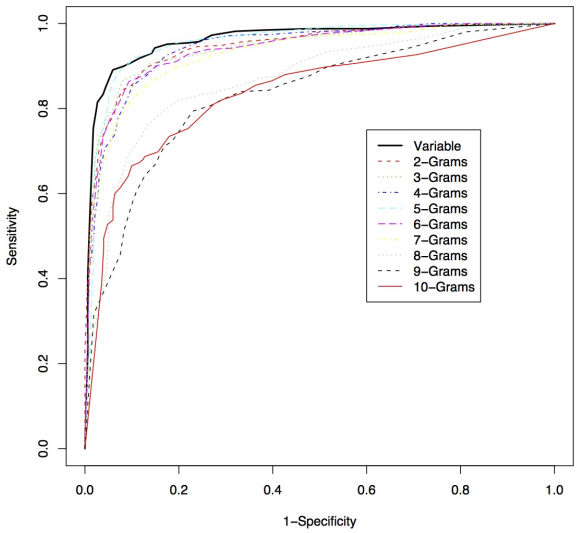


Fig. 9: ROC curves comparing bagging results for each n-gram size using PCA variable selection.

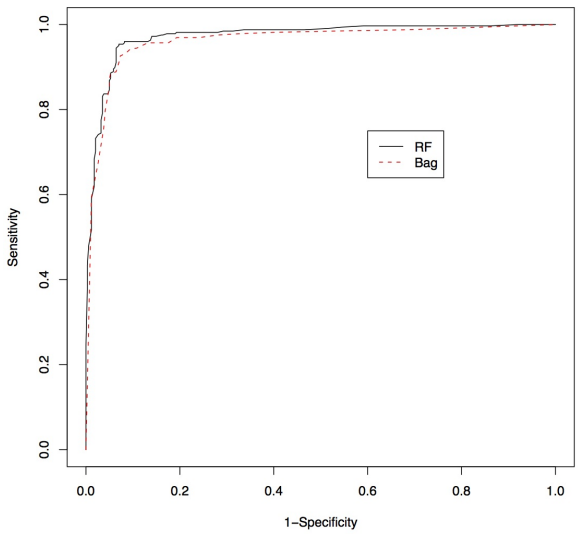


Fig. 10: ROC curves comparing random forest and bagging results using random forest variable selection.

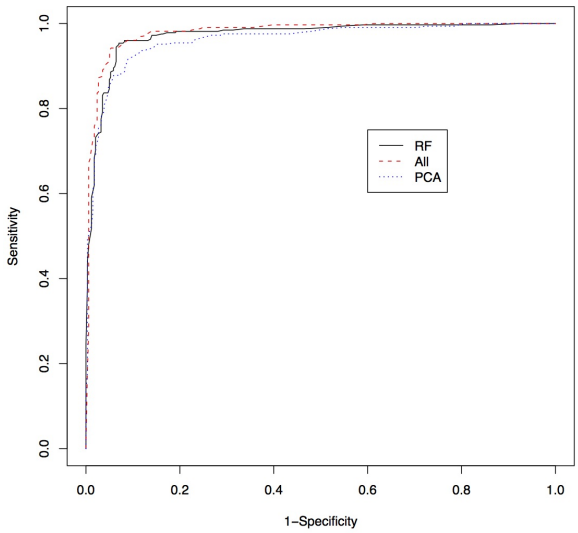


Fig. 11: ROC curves comparing the variable selection methods (all variables, random forest selection, PCA selection)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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, "Opem: 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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:

1. 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

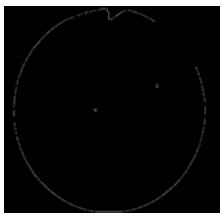
3. experimental results

The paper presents the effect of tuning canny parameters (threshold and sigma) to show how the performance of canny algorithms depends on these parameters.

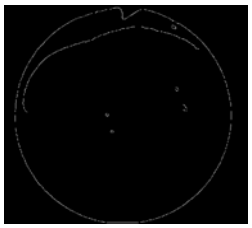
Canny Edge detector algorithm computes pixel's gradient values by using a Gaussian filter, for that two threshold are applied upper and lower, to detect strong and weak edges respectively, the ones between are added to edges only if they are connected to strong edges. The accuracy of this algorithm guarantees 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.

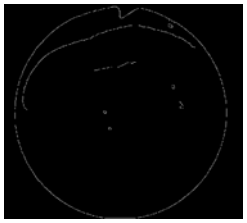
Original image



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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

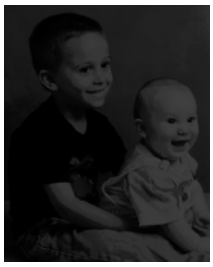

sigma=2.5 sigma=3


sigma=4 sigma=6


sigma=8 sigma=10


sigma=12 sigma=14

Fig.2 the impact of canny sigma with fix threshold=0.14 (for endoscopic image)




original image threshold=0.2


threshold=0.16 threshold=0.14

Fig.3. the impact of cany threshold with fix sigma=1.414 (for non medical image)


sigma=1.6 sigma=2


sigma=4 sigma=6

Fig.4. the impact of canny sigma with fix threshold=0.14 (for non medical 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 oc Computer and Information Technology of Hodeidah University, Hodedah, 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 - Algoriths 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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, Bihac, 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 number, investigator ID, Evidence number, SN, Time, Notes	MD5	User 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
DEMF	**DEMF, 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,**

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

		and Time, etc.		**Court order number, Legislation secured with AES256 encryption**

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.

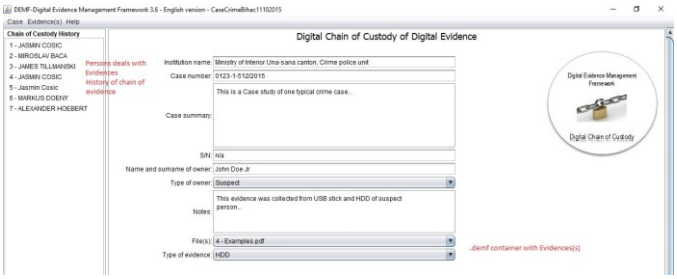


Figure 1 Example 1 of one real case

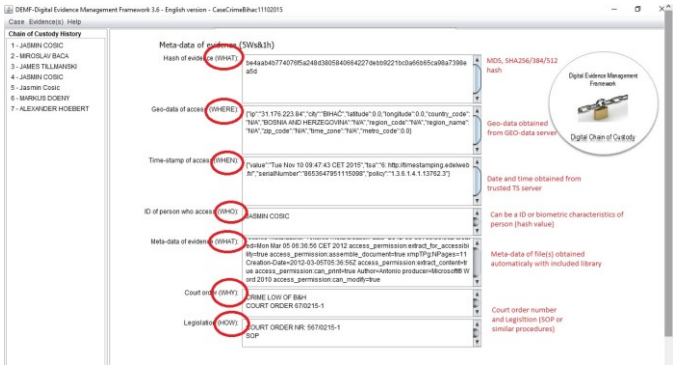


Figure 2 Example 2 of one real case

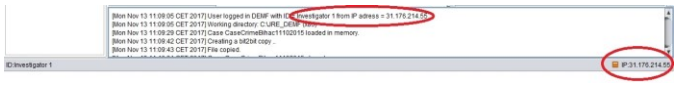


Figure 3 Example 3 of one real case with DEMF in action

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] *CaseDEMF_256K2017* created in memory.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[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 CDESF 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. Ćosić and M. Bača, "(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. Bača, "' 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 autonomics is the desired requirement of any CPS, so achieving autonomics is the necessary thing as nobody wants CPS to crash at any time. Hence autonomics 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, autonomics, 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 compos ability 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 compos ability 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 compos ability 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 autonomics 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 autonomics in CPS. It also provides some methods for self-healing approach. Ilsun et.al.[49] have presented a framework for autonomic

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 autonomics. 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 autonomics 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 autonomics 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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, Engergy, 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 Wirelesss 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 Appraoch", 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(ISCCCA), 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 Enggineering, 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 Handing 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 autonomics 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 Uncertainity 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_alsaif@hotmail.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].

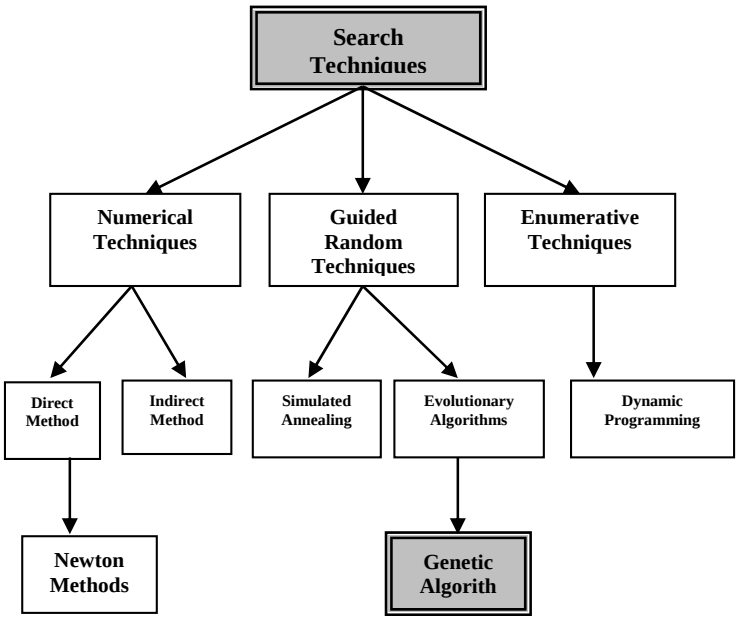


Figure (1): Optimization Techniques

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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]

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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]

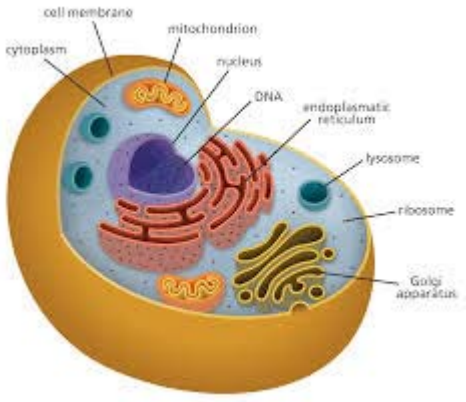


Figure (3) : the cell

- **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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

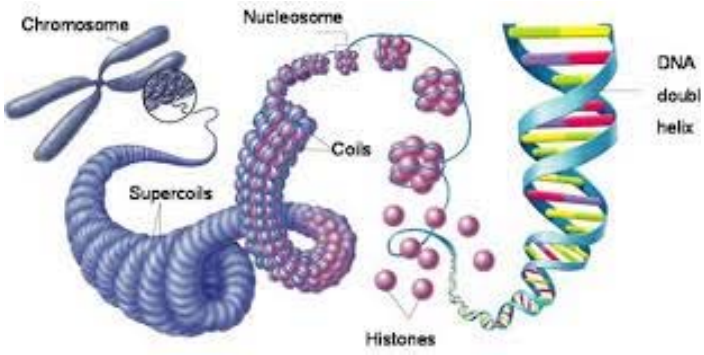


figure (4) : The Chromosome

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]

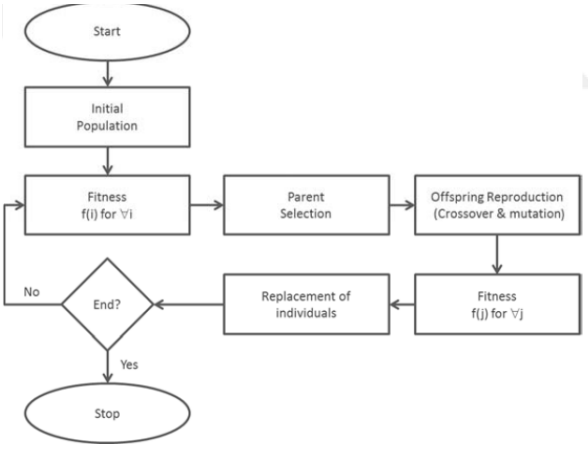


Figure (5). GA flowchart

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

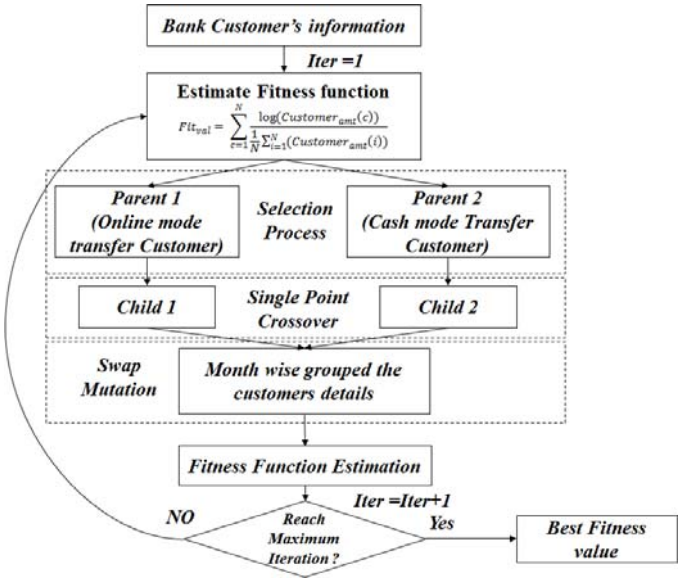
- Estimate the fitness function based on the transaction amount,
- After estimating the fitness function the parent chromosomes are generated at selection process.
- In the selection process clustered the customer by their mode of transaction.
- Next generated the child chromosomes from the above generated parent chromosome, this step is performed at the single point crossover operation.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- 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 $\text{Cluster}_{\text{Customer}}$

Output: Optimized value for Detection $\text{Optimum}_{\text{value}}$

Procedure:

Step1: Initialize maximum iteration $\text{Max}_{\text{iter}} = 10$

Step2: Estimate Fitness function based on the transaction amount,

$$\text{Fit}_{\text{val}} = \sum_{c=1}^{N} \frac{\log(\text{Customer}_{\text{amt}}(c))}{\frac{1}{N}\sum_{i=1}^{N}(\text{Customer}_{\text{amt}}(i))}$$

Where,

$\text{Customer}_{\text{amt}}$ – 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,

$$\text{CP} = \text{randint}(1,1,[1\ \text{length}(\text{Parent1})])$$

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,

$$\text{Month}_{\text{num}} = \text{Child}\ (\text{Date}_{\text{trans}})$$

Step6: Go to step2 for calculating the fitness value in monthly wise,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 :

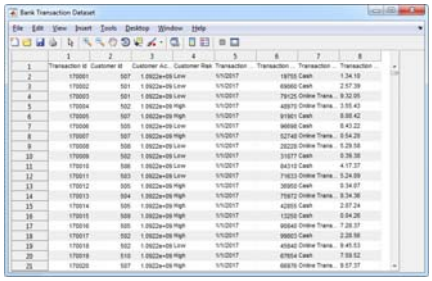


Figure 6: Output list of initial dataset

-The figure 7 shows the final transaction dataset:

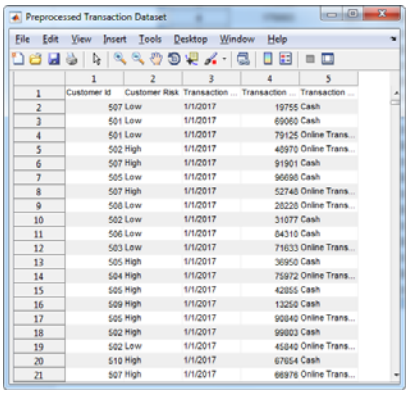


Figure 7: Final Transaction Dataset

-Next process is to cluster the customers report. The entire transaction details of each customer have been clustered as shown in Figure 8 below:

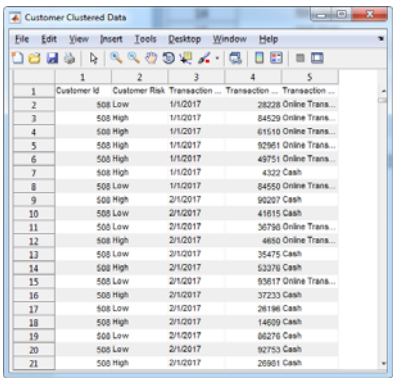


Figure 8: Final Transaction Dataset

-Figure 9 shows the result of this selection process:

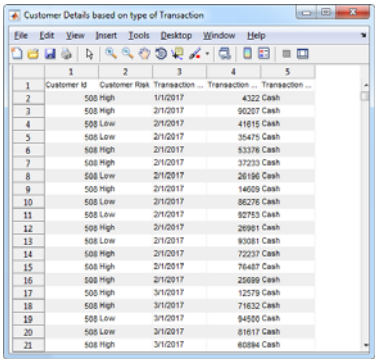


Figure 9: Customer details based on the selection process

-Figure 10 shows the details of mutation details.

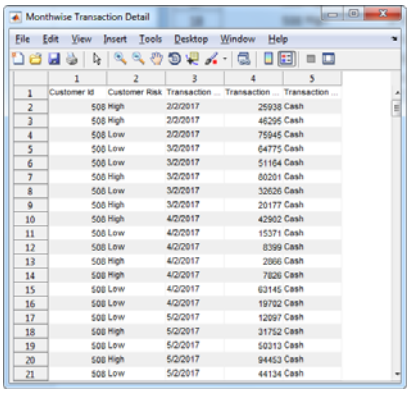
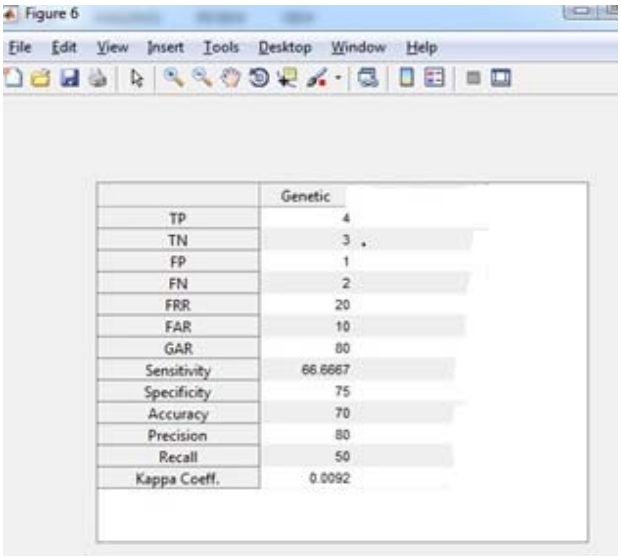


Figure 10: Month wise transaction details.

- Results of implementation of the proposed algorithm



	Genetic
TP	4
TN	3
FP	1
FN	2
FRR	20
FAR	10
GAR	80
Sensitivity	66.6667
Specificity	75
Accuracy	70
Precision	80
Recall	50
Kappa Coeff.	0.0092

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 11: Performance GA

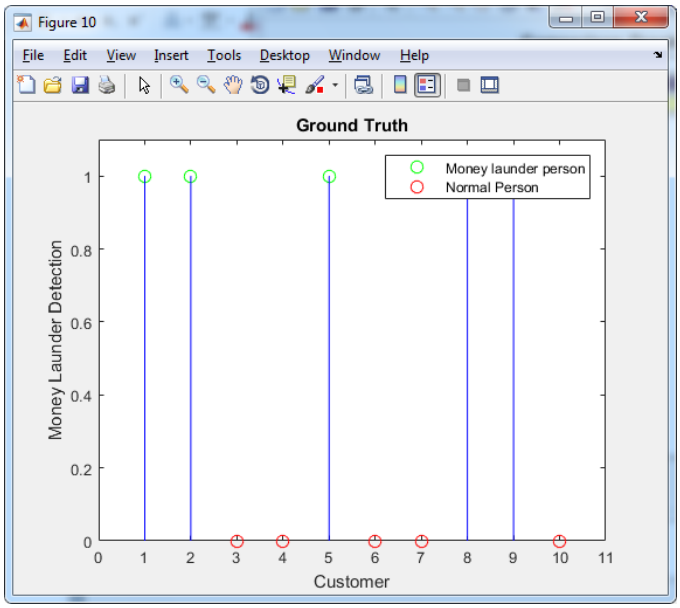


Figure 11: Comparison by ground truth

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 ystems", 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 forFinancial 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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(k
k)+1,:);
      x=x+1;
   end
end
function [Total_amt,cus_amt1]=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
```


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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; meandred 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].

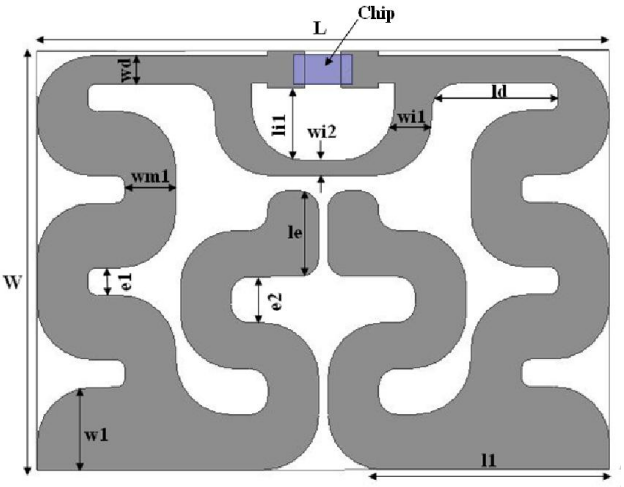
Distinct ultra-high frequency (UHF) band can be devoted according to the regulations in the work region. To cover the entire passive UHF band, the allocated frequency range must be between 860 to 960 MHz [1].

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].

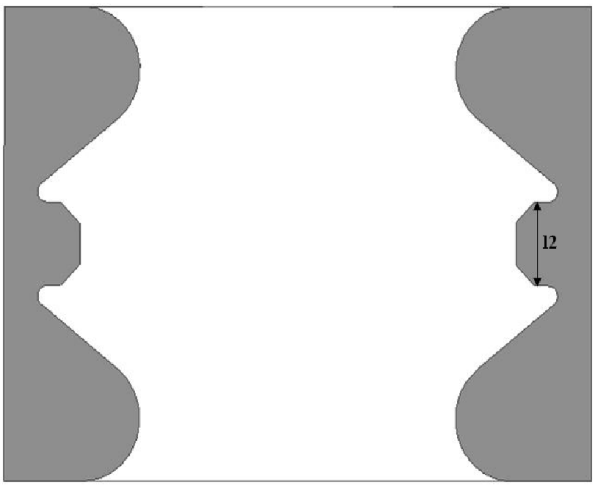
In this paper, a new compact symmetric passive UHF tag antenna operating at 915 MHz is presented. It is composed of a radiating modified meandered dipole strips, an altered T-matching and a modified ground plane. Each folded dipole arm is shorted to the suggested ground plane via shorting plate.

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 ε_r=2.2, loss tangent tang (δ) =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)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

(c)

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 (wi1 or wi2 or lp or wp) while others stay unaltered (L=31mm,W=22.75mm,wd=1.5mm,ld=6.4357mm, li1=3.9625mm,wm1=2.75mm,l1=12.25mm,le=4.65mm, w1=4.5mm, e1=1.5mm, e2=2.5mm, l2=4mm).

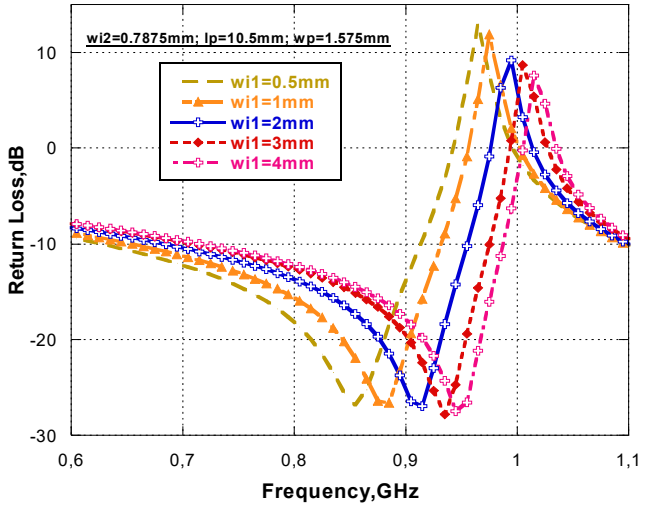


Figure 2. Variation of the return loss for different wi1 values.

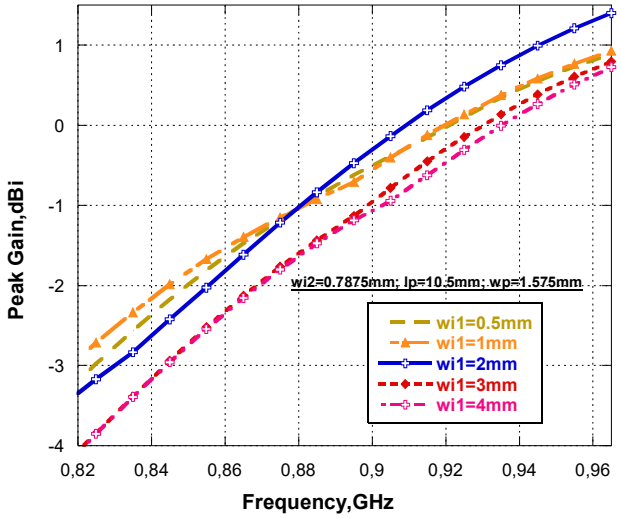


Figure 3. Variation of the peak gain for different wi1 values.

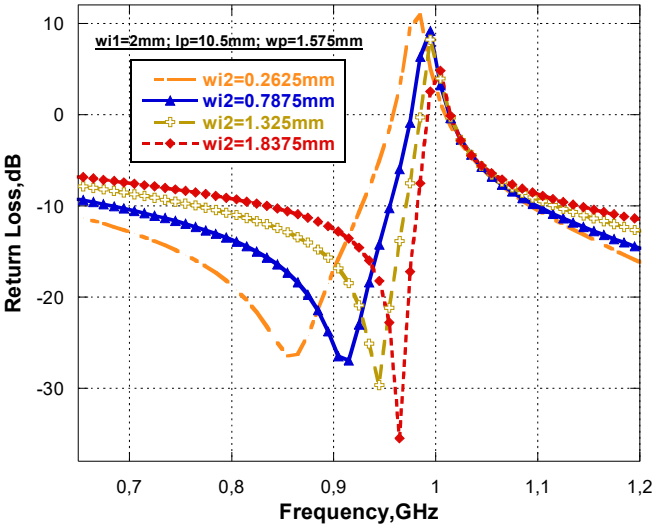


Figure 4. Variation of the return loss for different wi2 values.

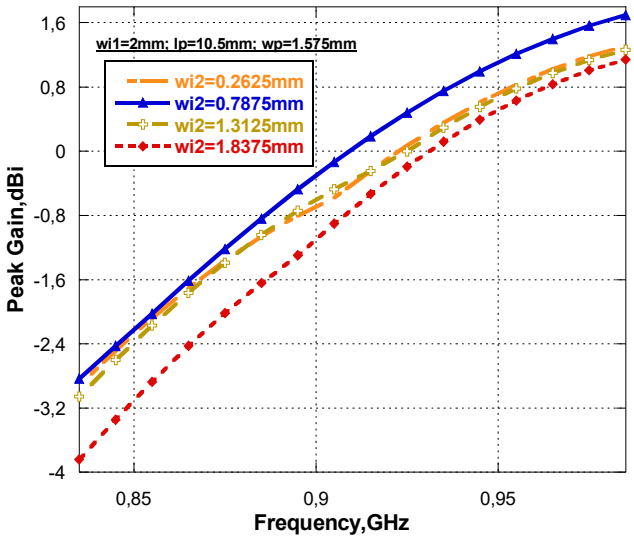


Figure 5. Variation of the peak gain for different wi2 values.

The dimensions of the width wi1 and wi2 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 wi1 or wi2 increases both the central frequency and its peak gain.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

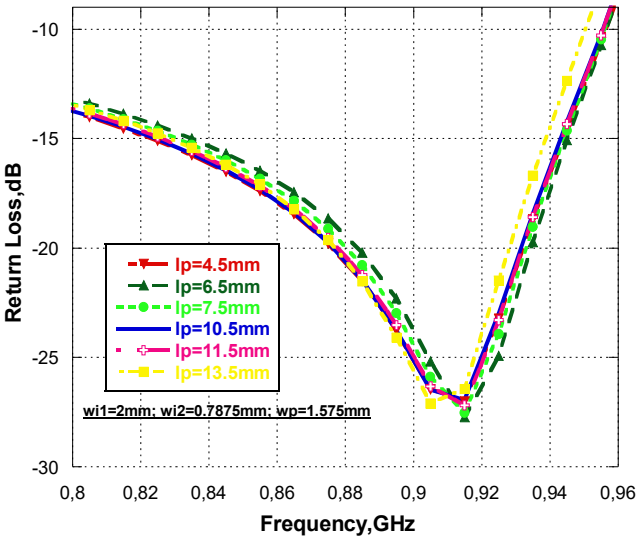


Figure 6. Variation of the return loss for different lp values.

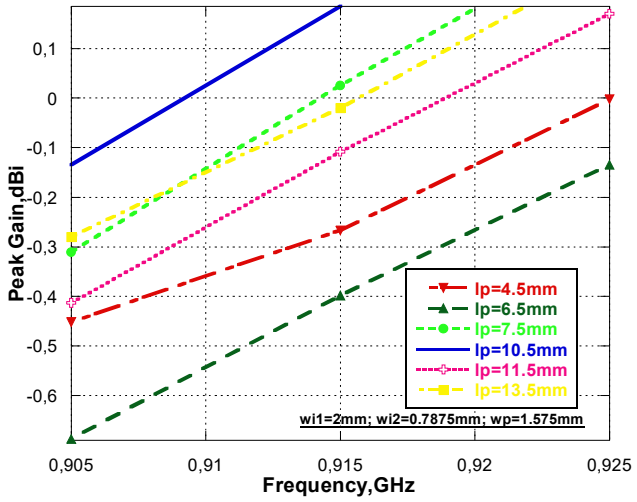


Figure 7. Variation of the peak gain for different lp values.

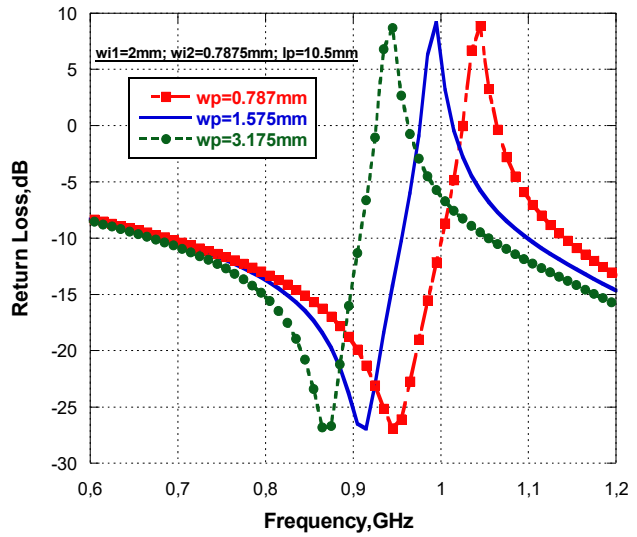


Figure 8. Variation of the return loss for different wp values.

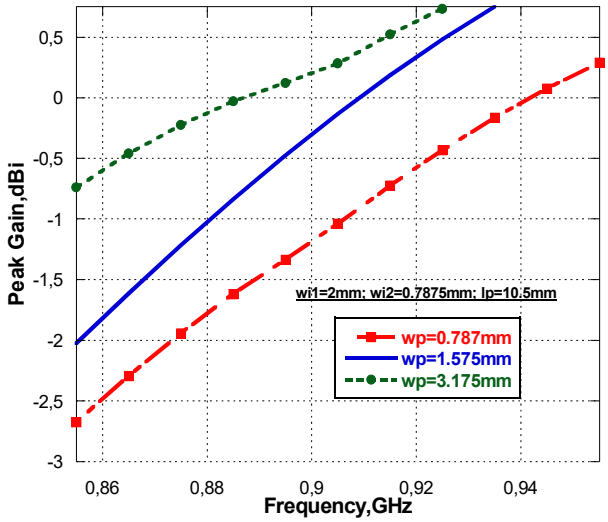


Figure 9. Variation of the peak gain for different wp values.

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

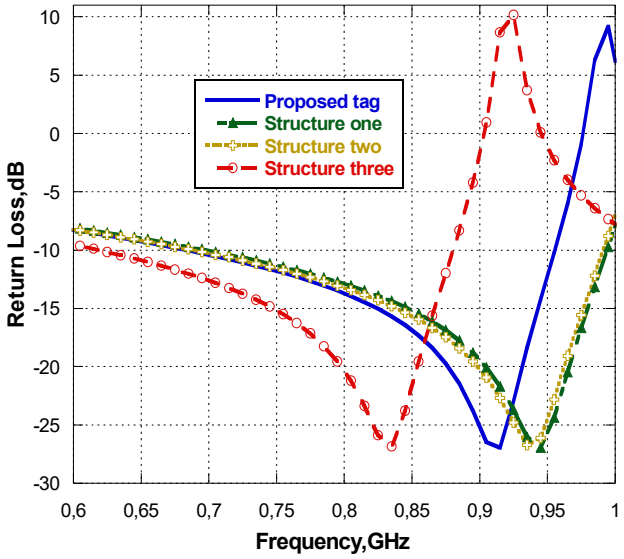


Figure 10. Return loss behavior comparison between the proposed tag and three other different structures.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

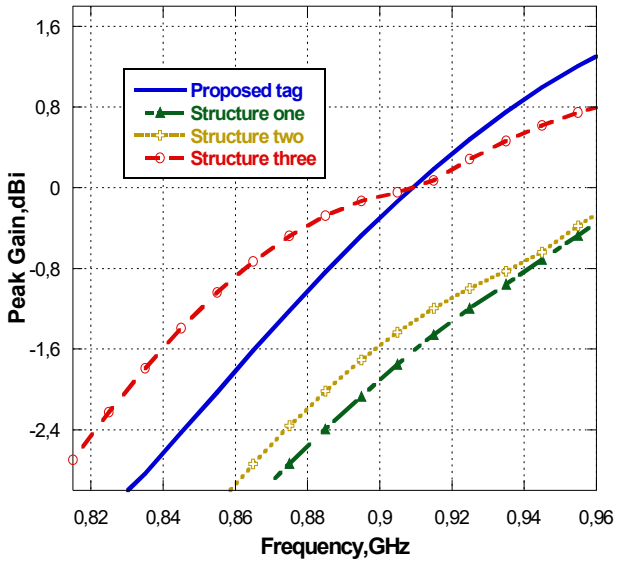


Figure 11. Peak gain behavior comparison between the proposed tag and three other different structures.

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.

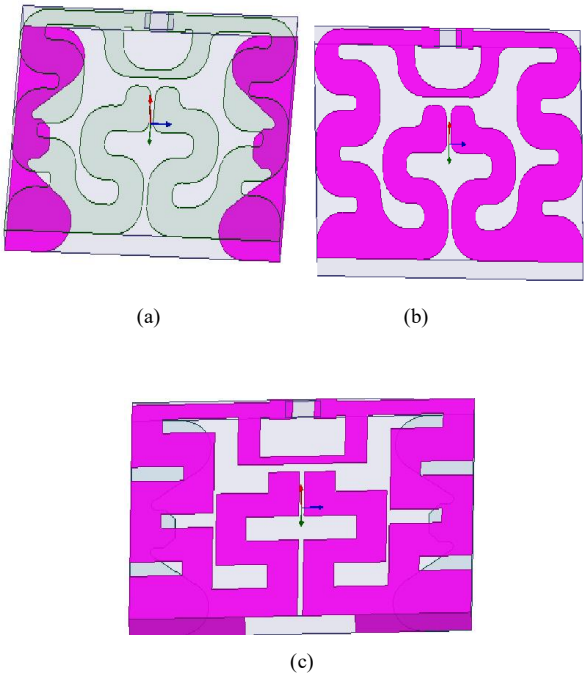


Figure 12. Three other different structures: (a) Structure one, (b) Structure two and (c) Structure three.

- 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 1.01dBi.
- 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.

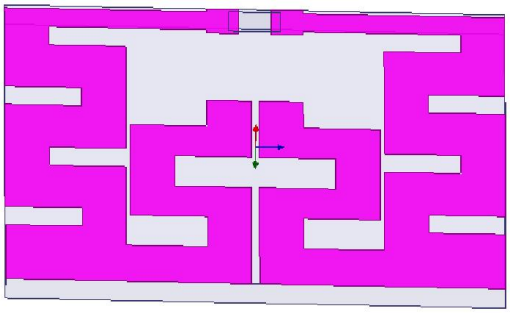


Figure 13. The conventional meandred dipole arms structure.

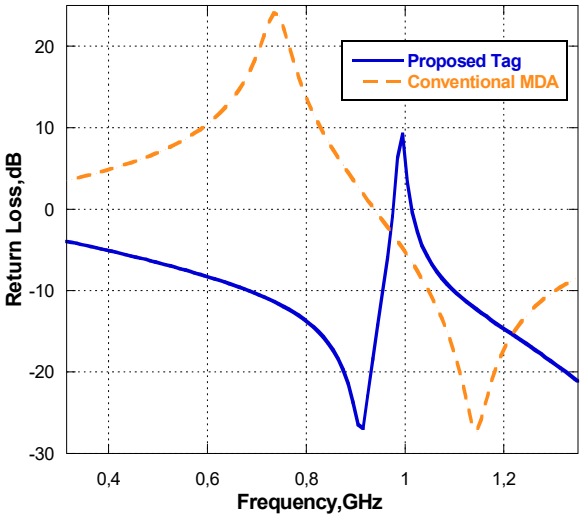


Figure 14. Return loss behavior comparison between the proposed tag and the conventional meandred dipole arms structure.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

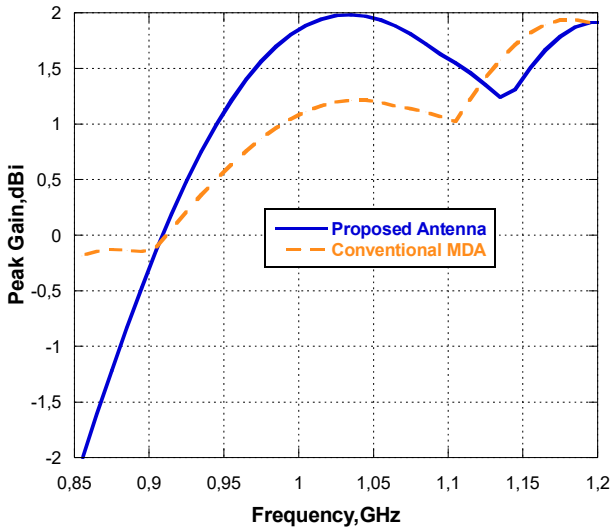


Figure 15. Peak gain behavior comparison between the proposed tag and the conventional meandred dipole arms structure.

- 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)
L	*31*	*l1*	*12.25*
W	*22.75*	*le*	*4.65*
wd	*1.5*	*w1*	*4.5*
ld	*6.4357*	*e1*	*1.5*
wi1	*2*	*e2*	*2.5*
wi2	*0.7875*	*l2*	*4*
li1	*3.9625*	*wp*	*1.575*
wm1	*2.75*	*lp*	*10.5*

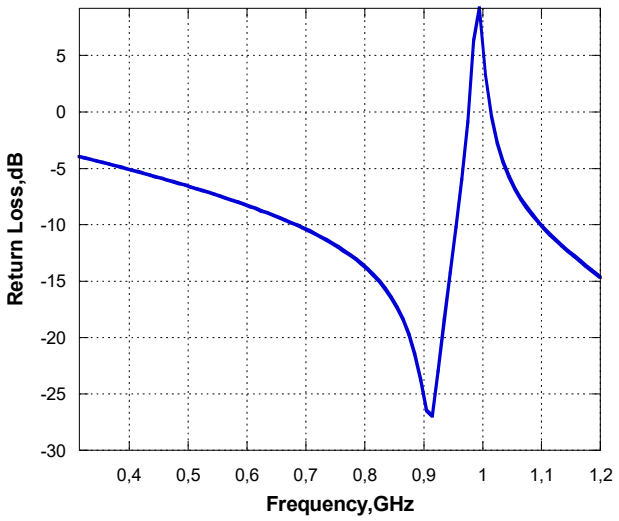


Figure 16. HFSS simulated return loss of the optimised UHF Tag.

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.

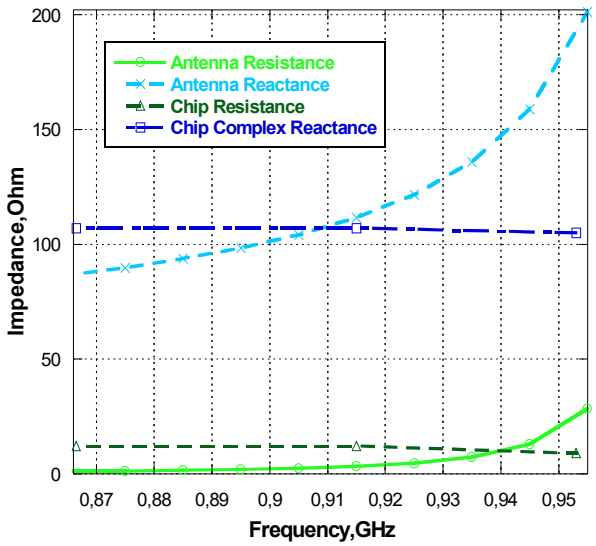


Figure 17. Input impedances in a region of interest.

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) Ω 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) Ω at the same frequency.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

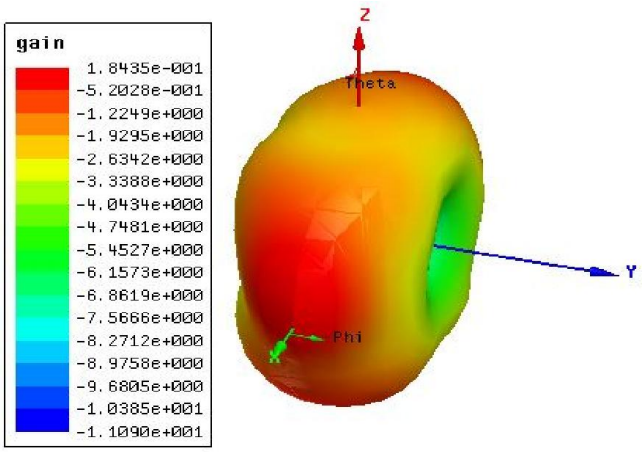


Figure 18. Maximum 3-D gain of the optimised UHF Tag at 915MHz.

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.

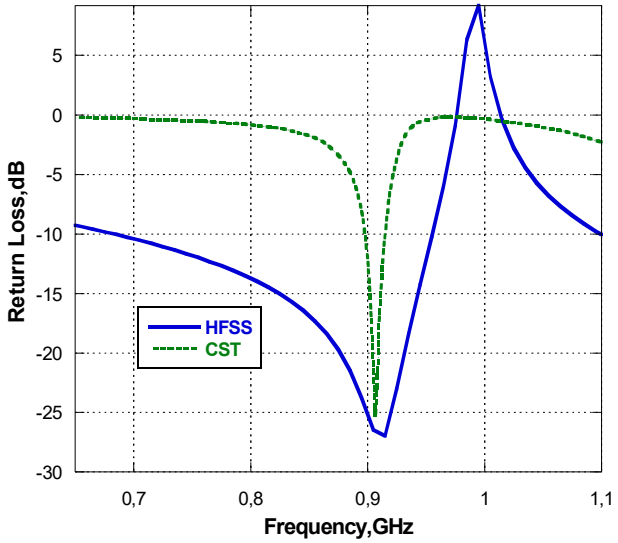


Figure 19. HFSS and CSTsimulated return loss comparison.

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. PROPESED 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 Miniaturezed 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

Dr.P.Sujatha
Associate Professor
Department of Computer Science
Vels University
Chennai, India
suja.research@gmail.com

***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 explicitely 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 separeated 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 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 theses 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

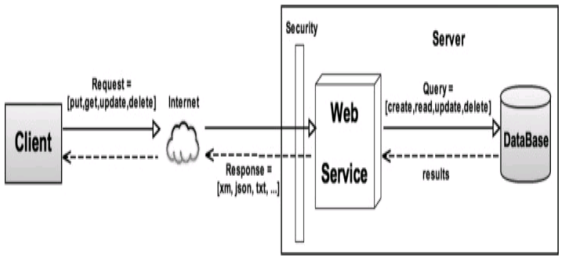


Fig 2. Secured communication flow between Client, Web service and Database.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

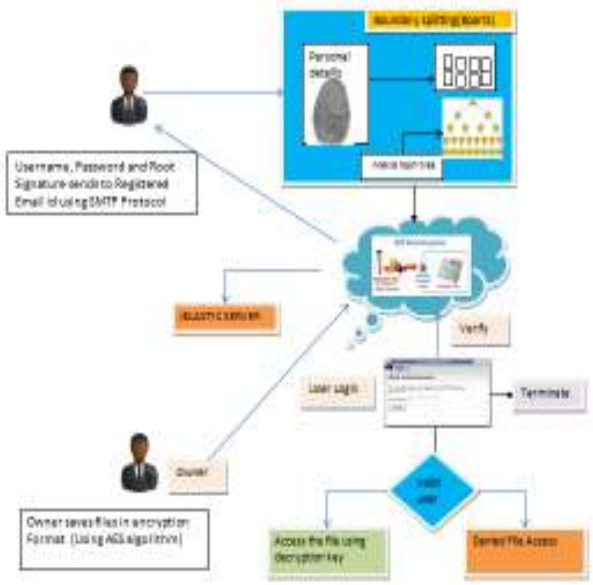


Fig.3 Architecture diagram on the authentication process of Jelastic Server

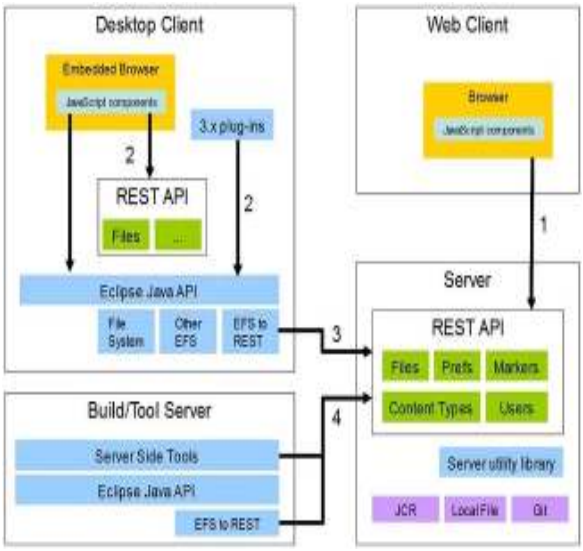


Fig 4. Layers of desktop, Server and Web-Client

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

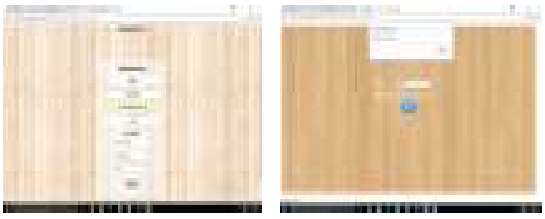


Fig 5. Registration and authentication check of a user

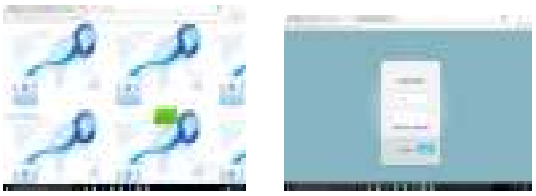


Fig 6. Login form of a user's screen

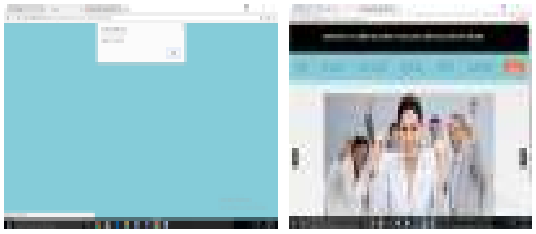


Fig 7. Login form of a user and home page

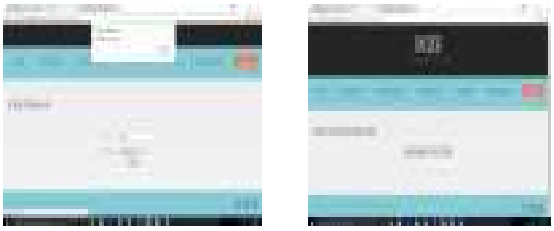


Fig 8. File upload and its respective details

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

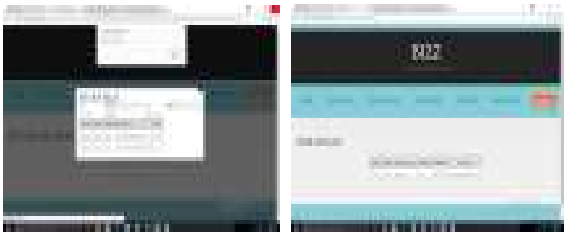


Fig 9. File sharing and validity details

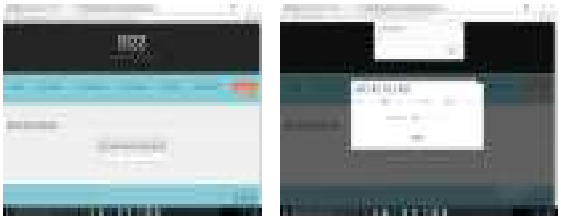


Fig 10. File receipt for the user

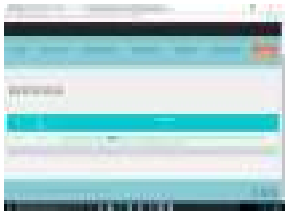


Fig 11. File access through secure key and receipt of files

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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/ijsia.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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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
SELECT count(*) as c FROM store_sales, household_demographics ,time_dim, store WHERE store_sales.ss_sold_time_sk = time_dim.t_time_sk and store_sales.ss_hdemo_sk = household_demographics.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_demographics.hd_dep_count = 5 and store.s_store_name = 'ese' order by c limit 100;	SELECT BITMAP_CARDINALITY(COMPOUND_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_proj__ WHERE t_minute_greater_equal_30=TRUE UNION ALL SELECT `_bitmaps` AS bitmaps FROM tpcds__orc_1__table96_hd_dep_count_proj__ WHERE hd_dep_count=5 UNION ALL SELECT `_bitmaps` AS bitmaps FROM tpcds__orc_1__table96_s_store_name_proj__ WHERE s_store_name = 'ese')d order by e limit 100;

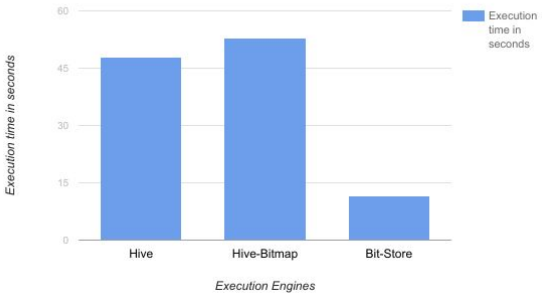


Fig. 1. Execution times for TPC-DS query 96

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.

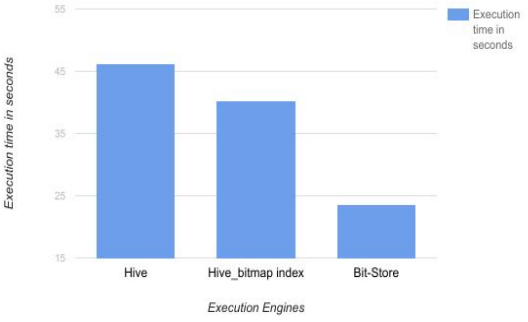


Fig. 2. Execution times for TPC-DS query 42

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE III. QUERY 42 - ORIGINAL AND RE-WRITTEN QUERIES

Original Query	Rewritten Query
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 ;	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) d)e, (SELECT DISTINCT `_bucketname` AS bucketname FROM tpcds__orc_1__table42_d_year_proj__)b GROUP BY bucketname, offset;

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.

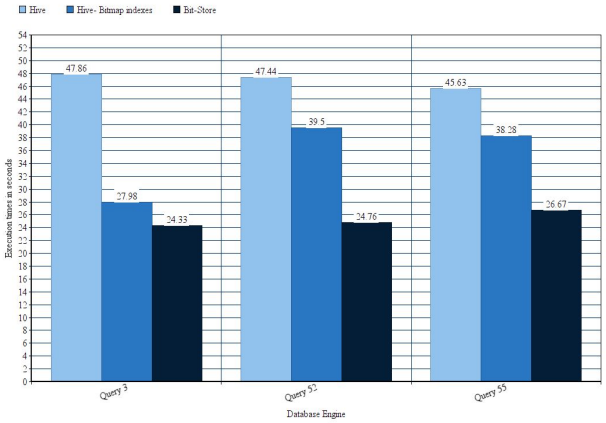


Fig. 3. Execution times for TPC-DS query 3, 52 and 55

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
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 ;	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_TERNARY(bitmap1.bitmap, bitmap3.bitmap, bitmap2.bitmap), bitmap4.bitmap)) FROM bitmap1, bitmap2, bitmap3, bitmap4

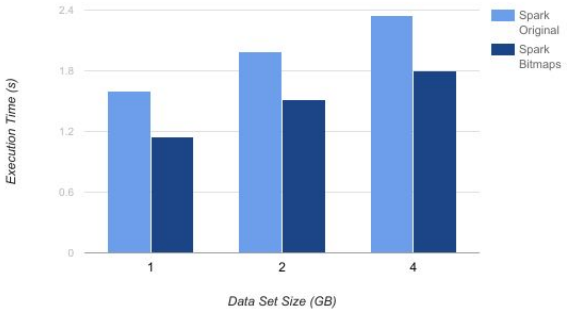


Fig. 4. Spark SQL TPCDS Query 96

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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. ICCAE '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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 (Lukas-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 Kanede 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 nxn 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:

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t). \quad (1)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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:

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) + \partial I \partial x \delta x + \partial I \partial y \delta y + \partial I \partial t \delta t + H.O.T. \quad (2)$$

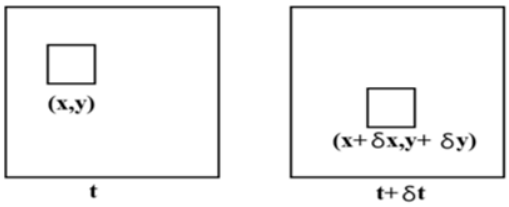


Figure 1 *image at point (x,y,t) is same as at point (x + δx,y + δy,t + δt)*

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:

$\partial I\ \partial x\ \delta x + \partial I\ \partial y\ \delta y + \partial I\ \partial t\ \delta t = 0$ *or* $\partial I\ \partial x\ \delta x\ \delta t + \partial I\ \partial y\ \delta y\ \delta t + \partial I\ \partial t\ \delta t\ \delta t\ |\{z\} = 1 = 0$ and finally :

$\partial I\ \partial x\ vx + \partial I\ \partial y\ vy + \partial I\ \partial t = 0.$ (3)

Here $vx = \delta x\ \delta t$ and $vy = \delta y\ \delta t$ are the x and y components of image velocity or optical flow and ∂I ∂x, ∂I ∂y and ∂I ∂t are image intensity derivatives at (x, y, t). We normally write these partial derivatives as:

$Ix = \partial I \partial x, Iy = \partial I \partial y$ *and* $It = \partial I \partial t$ (4)

Where $\nabla I = (Ix, Iy)$ is the spatial intensity gradient and $\vec{v} = (vx, vy)$ is the image velocity or optical flow at pixel (x, y) at time t. $\nabla I.\vec{v} = -It$ 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 *vx,vy* differentials are x and y optical flow components and *(Ix,Iy,It)* which are intensity derivatives. This equation can be rewritten more compactly as:

$$(Ix, Iy).(vx, vy) = -It \quad (5)$$

or as:

$$(\nabla I.\vec{v}) = -It \quad (6)$$

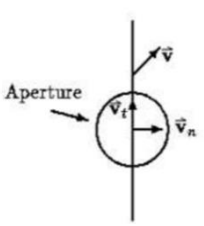


Figure 2(a) Aperture

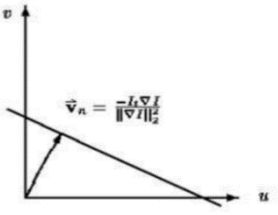


Figure 3(b) Normal Velocity

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 Constrain Equation produces a line in $\vec{v}$ = (vx,vy) sepctrum. One speed on this line is the right speed. The speed with the littlest extent on that line is the ordinary speed vn̂. The greatness and bearing of the typical speed, , $\vec{v}$n = vnn̂ 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. :

$\overrightarrow{v_n} = v_n.\hat{n} = \frac{-It(Ix,Iy)}{||\nabla I||2^2}$. (8) is the raw ordinary speed and ∇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:

$$\overrightarrow{v_n}.\hat{n} = v_n \quad (9)$$

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

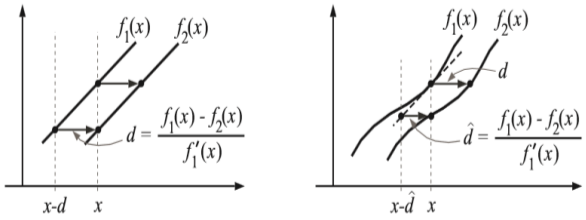


Figure 2 Gradient constraints on Linear and Non-Linear signals

displacement of a linear signal (left), the difference in signal values at a point divided by the slope gives the displacement. For

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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:

$$I(\vec{x}, \mathrm{t}) = I(\vec{x}, +\vec{u}, \mathrm{t}+1) \quad (1.1)$$

Where $I(\vec{x}, \mathrm{t})$ is image intensity as a function of space $\vec{x} = (x, y)^T$ and time t, and $\overrightarrow{\vec{u} = (u_1, u_2)^T}$ in the 2D velocity. To derived an estimator for 2D velocity $\vec{u}$, 1D case was considered first. Let f1(x) and then f2(x) assigned 1D signals (images) at two time instants. As shown in Figure 3, assuming that f2(x) be a interpreted version of f1(x); i.e., let $f2(x)=f_1(x-d)$ where d means translation. A Taylor series expansion of $f_1(x-d)$ about x is given by

$$f_1(x-d)=f_1(x)-df_1'(x)+O(d2f_1'') \quad (1.2)$$

where $f' \equiv df(x)/dx$. This expansion can be rewritten as the difference between the two signals at location x as $f1(x)-f2(x)=df_1'(x)+O(d2f_1'')$. Ignoring second- and higher-order terms, An approximation can be obtained to d:

$$d = \frac{f1(x)-f2(x)}{f_1'(x)}. \quad (1.3)$$

The 1D case simplifies to 2D. As above, assuming the displaced image is approximated very well by a first-order Taylor series:

$$I(\vec{x}, +\vec{u}, t+1) \approx I(\vec{x}, t) + \vec{u}.\nabla I(\vec{x}, t) + I_t(\vec{x}, t) \quad (1.4)$$

Where $\nabla I \equiv (Ix, Iy)$ *and* I_t denotes spatial and temporal partial derivatives of the image I, and $\overrightarrow{\vec{u} = (u_1, u_2)^T}$ 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 (Tc) 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

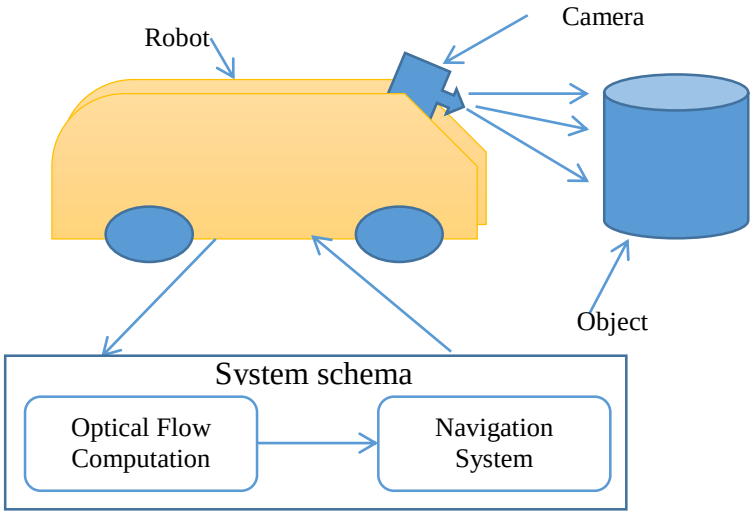


Figure 4 ***Optical Flow navigation System***

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,*

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

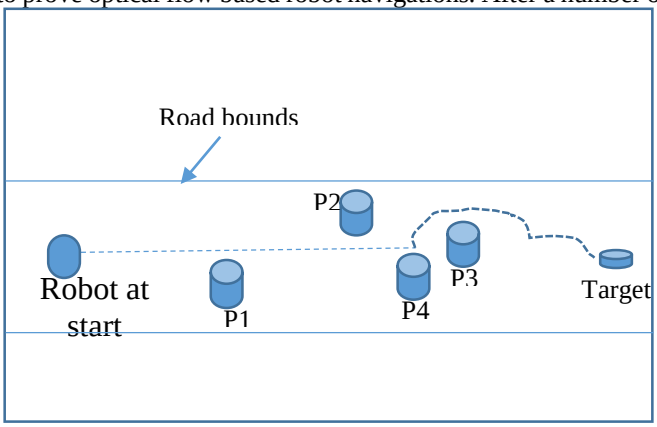


Figure 5 Navigation Environment

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 Krajník [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 sonsors 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 enviroment (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 exepensive 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[1] J. Ng, "An Analysis of Mobile Robot Navigation Algorithmsi 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[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. Bouthemy and C. Kervrann, "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. Krajník, 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 Technlology (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 pursing 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Dataset of Graphs and Sub-graphs: Storage Representation and its Graphical form

Rachna Somkunwar[1] and Dr. Vinod M. Vaze[2]

[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 combinatory 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].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

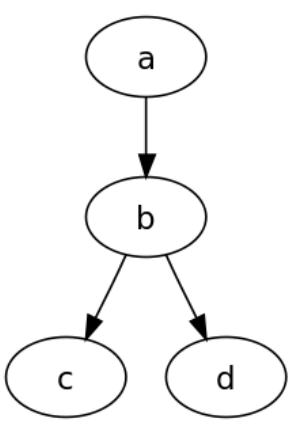


Figure.1. Directed Graph

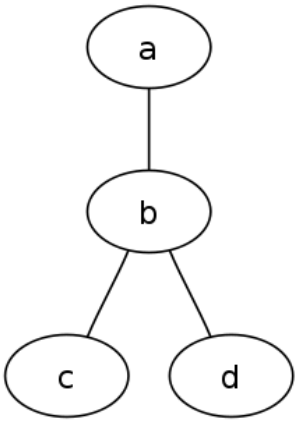


Figure. 2. Undirected Graph

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Fixed valance graph, where the number of vertices is equal to the number of links.

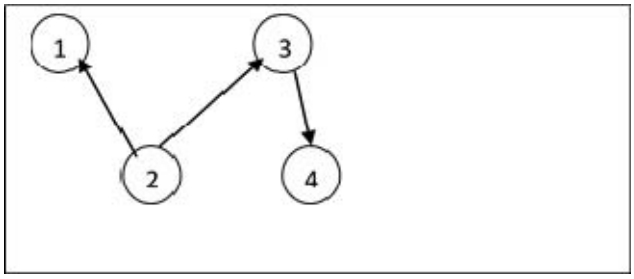


Figure. 3. Random Graph

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 aij= 1vertex i is neighbor of vertex j.
else aij=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 *V| 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

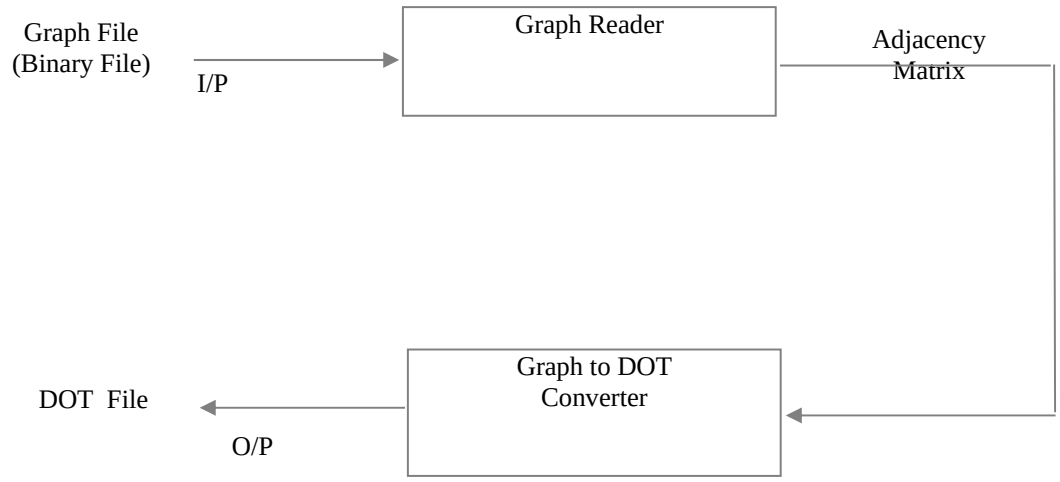


Figure 4. Block diagram to convert Binary file into DOT file

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>

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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]

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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]
}

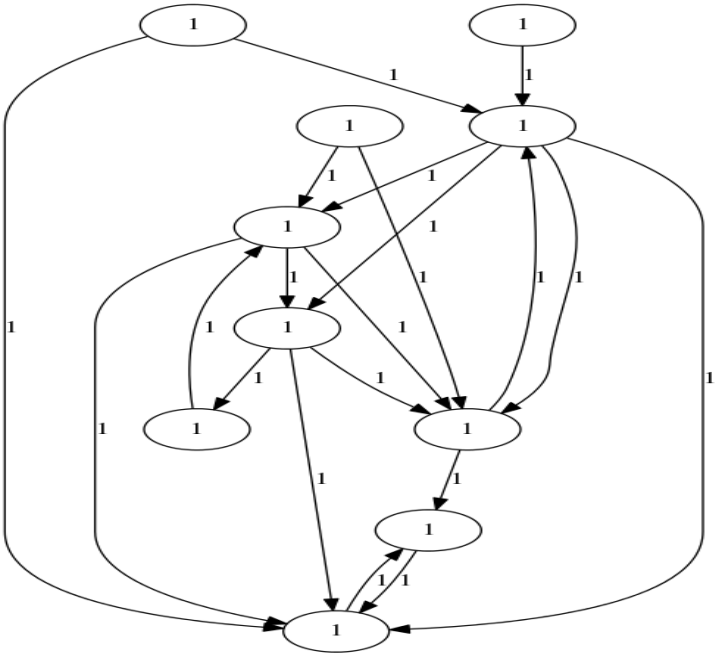


Figure 5. Graphical representation of Gen graphs with 10 nodes and 20 edges

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 Intractibility: 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 distinguishment 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 Retrilval, 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.

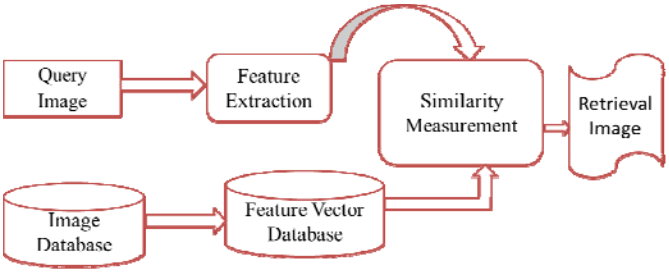


Figure 1: Basic CBIR Flow

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 Fingureprint Image Retrival

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

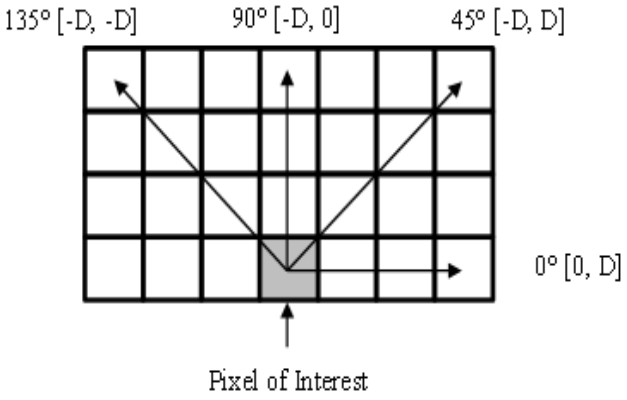


Figure 2: Directional Analysis for GLCM (P_0, P_{45}, P_{90} & P_{135})

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	$\mathbf{F_1} = \sum_i \sum_j \{p(i,j)\}^2$
2.	Contrast	$\mathbf{F_2} = \sum_{n=0}^{Ng-1} n^2 \{\sum_i \sum_j p(i,j)\}$
3.	Entropy	$\mathbf{F_3} = \sum_i \sum_j p(i,j) \log(p(i,j))$
4.	Inverse Difference	$\mathbf{F_4} = \sum_i \sum_j \frac{p(i,j)}{1+(i-j)^2}$

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.

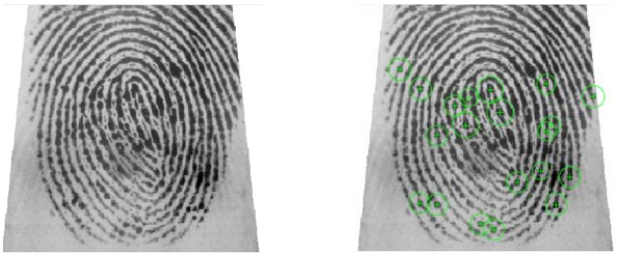


Figure 3: Detection of SURF feature points (a) Input Image (b) Interest Points

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 Wavelate

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 ∂x and ∂y of the Gaussian envelop as follows [10].

$$\psi(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{[1-(\frac{n}{2})(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2})+2\pi j\omega x]} \quad (1)$$

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-m(xcosθ + ysinθ) , y=a-m(-xcosθ + ysinθ), θ= nπ / K, m={0,…,S-1} and n= {0,…., 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	• Small Dimension Feature Vector. • Covers the Direction Property • Less Computation Time • Multiorientation & Robust	• Only consider Gray scale images.
SURF [8]	• Invariant to Rotation and Scale detector and descriptor. • Images under different conditions such as low light images, blurred images are also recognized.	• -Only consider Gray scale images. • High Dimension Feature Vector.
GBW [10][11]	• Extracted impo- rtant components of images • Invariance to illumination, rotation, scale, transform	• 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]:

$$\Delta d = \sqrt{\sum_{i=1}^{n}(|Q_i - D_i|)^2} \quad (4)$$

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)

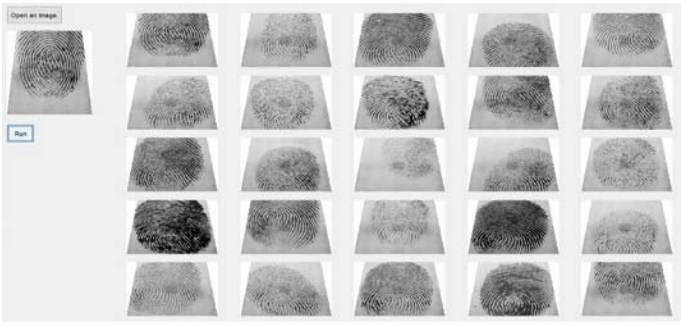


Figure 4: Retrieval Result

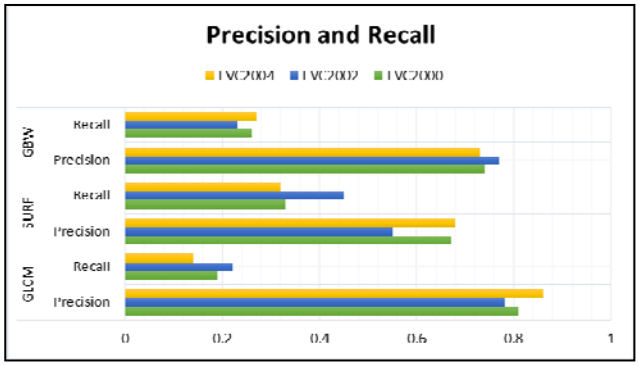


Figure 5: Precision and Recall

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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] Pönitz, 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] Khatere 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

PROACTIVE DETECTION OF CATASTROPHE TRENDS FOR RESCHEDULING REAL-TIME SYSTEMS WITH SCENARIO SHIFT

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.

[1]christypersya@gmail.com,[2]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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 cause identified for the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 arbitrative 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 [C^2MDS]

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

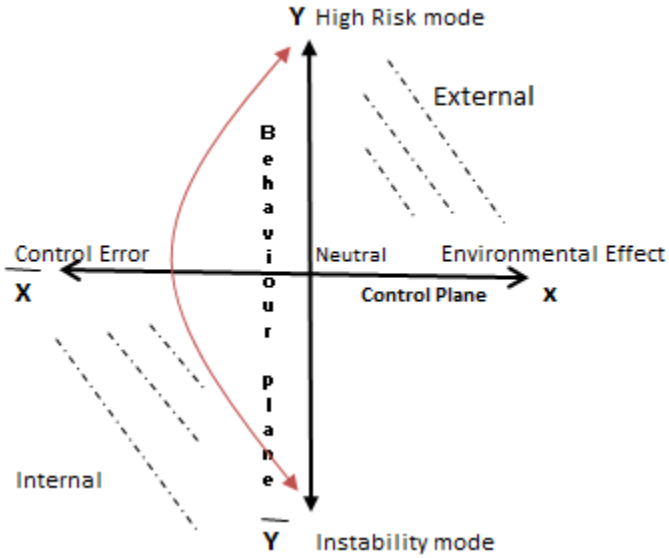


Fig. 1. Linear model in RTS

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

risk mode leading to oscillations between them. In this situation, the linear model is insufficient to depict the transition along $Y\bar{Y}$ 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.

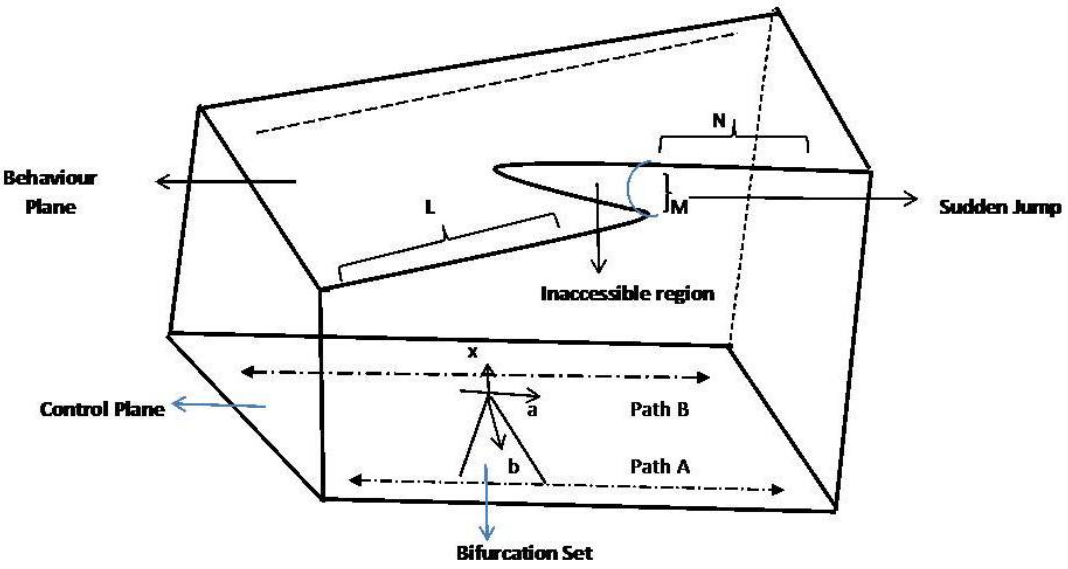


Fig. 2. The cusp model in RTS.

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 inflexion) 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

$$V_{ab}(x) = \frac{1}{4}x^4 + \frac{1}{2}ax^2 + bx \qquad (3.1)$$

The balanced surface of $V_{ab}(x)$ is

$$x^3 + ax + b = 0 \qquad (3.2)$$

The singular point of $V_{ab}(x)$ is

$$3x^2 + a = 0 \qquad (3.3)$$

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

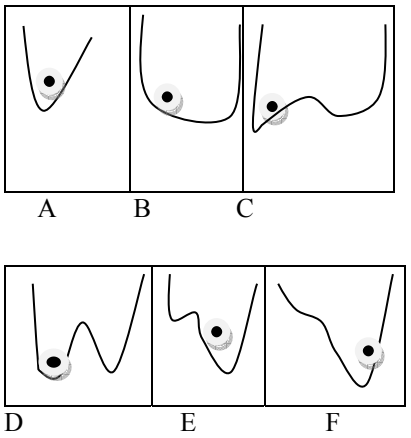


Fig. 3. Continuous increase leads to sudden shift

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.

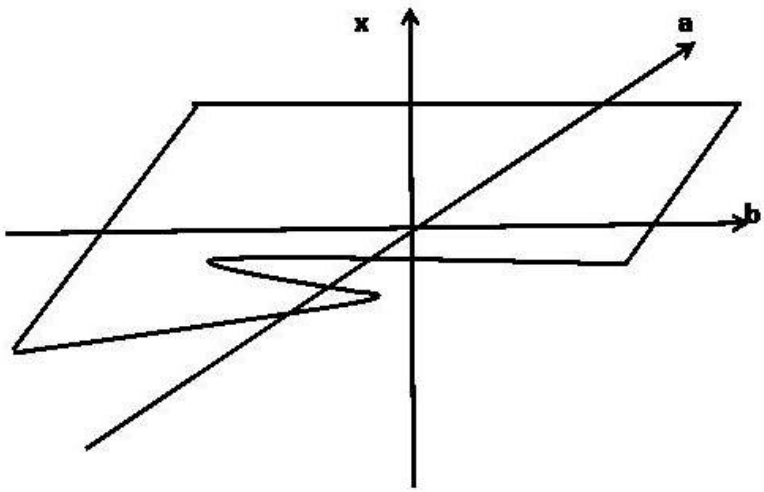


Fig. 4. Cusp model

The two different behaviors of the quartic function are separated by the curve shown in Fig 4.

$$8a^3 + 27b^2 = 0 \quad (3.4)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

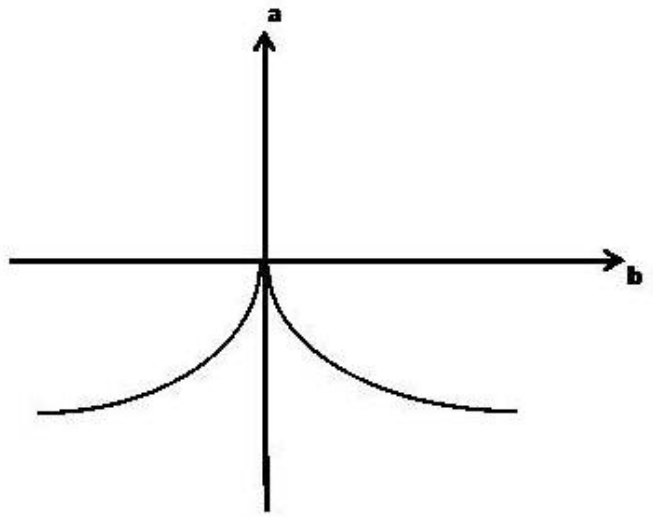


Fig. 5. Cusp Point

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 $S_1,S_2,\ldots..,S_n$ 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

service of the queue. There can be n queues uniquely coupled to a subsystem S_1,S_2, …..,S_n or one queue can cater more than one subsystem.

Whenever the underperformance of $S_1,S_2,…..,S_n$ 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

$$U = \sum_{i=1}^{n} \frac{C_i}{T_i} \leq 1 \qquad (3.5)$$

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 $\tilde{q}$.
ii) The queue length as q_{len}.
iii) The mean arrival rate is $\frac{1}{\lambda}$.
iv) The mean service rate is $\frac{1}{\mu}$.
v) The performance measure (ρ) of a uniprocessor queue is $\rho = \frac{\lambda}{\mu}$.
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 ($q_{len} \geq \tilde{q}$). 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.

$$\alpha(x) = \frac{N - x}{T} \tag{3.6}$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

$q_{en} \geq \tilde{q}$ indicates that for few cycles, few underperformance happening with the subsystem.
Therefore, P [underperformance happening] depends on P[$q_{en} \geq \tilde{q}$].

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., $q_{en} = q_{sc} + q_{pc}$.

$$P[q_{sc} + q_{pc} \geq \tilde{q}] = \rho^{(q_{sc}+q_{pc})} \quad \text{if } \rho < 1 \tag{3.7}$$

Because in M/M/1 model,

$$P_n = (1 - \rho)\rho^n, \quad n=0,1,2 \tag{3.8}$$

$$P[q_{sc} + q_{pc} \geq \tilde{q}] = 1 \quad \text{if } \rho \geq 1 \tag{3.9}$$

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., $\lambda = \mu.x$

Hence $\dfrac{\lambda}{\mu.x} = 1$

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.

$$\beta(x) = \frac{x}{T} \cdot \left[\frac{\lambda}{\mu.x}\right]^{q_{sc}} \cdot \left[\frac{\lambda}{\mu.x}\right]^{q_{pc}} \quad \text{if } \rho < 1 \tag{3.10}$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$\beta(x) = \frac{x}{T}.1 \quad \text{if } \rho \geq 1 \tag{3.11}$$

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,

$$drift(x) = \alpha(x) - \beta(x) \tag{3.12}$$

Substituting $\alpha(x)$ and $\beta(x)$ in the drift function,

$$= \frac{N-x}{T} - \frac{x}{T}\left(\frac{\lambda}{\mu.x}\right)^{q_{sc}} \cdot \left(\frac{\lambda}{\mu.x}\right)^{q_{pc}} \quad \text{if } \rho < 1 \tag{3.13}$$

$$= \frac{N-x}{T} - \frac{x}{T} = \frac{N-2x}{T} \quad \text{if } \rho \geq 1 \tag{3.14}$$

The potential function of the drift equation while $\rho < 1$ and $\rho \geq 1$ is given below:

$$V(x) = \frac{1}{T}\left\{Nx - \frac{1}{2}x^2 - \left(\frac{\lambda}{\mu}\right)^{q_{sc}+q_{pc}}\left(\frac{x^{1-(q_{sc}+q_{pc})}}{1-(q_{sc}+q_{pc})}\right)\right\} + c \quad \text{if } \frac{\lambda}{\mu.x} < 1 \tag{3.15}$$

$$V(x) = \frac{1}{T}\{N.x - x^2\} + c \quad \text{if } \frac{\lambda}{\mu.x} \geq 1 \tag{3.16}$$

To have the continuity of V(x), the potential function becomes

$$V(x) = \frac{1}{T}\left\{2Nx^2 - \frac{3}{2}x^4 - \left(\frac{\lambda}{\mu}\right)^{q_{sc}+q_{pc}}\left(\frac{1}{1-(q_{sc}+q_{pc})}\right)x^{1-(q_{sc}+q_{pc})}\right\} + c \tag{3.17}$$

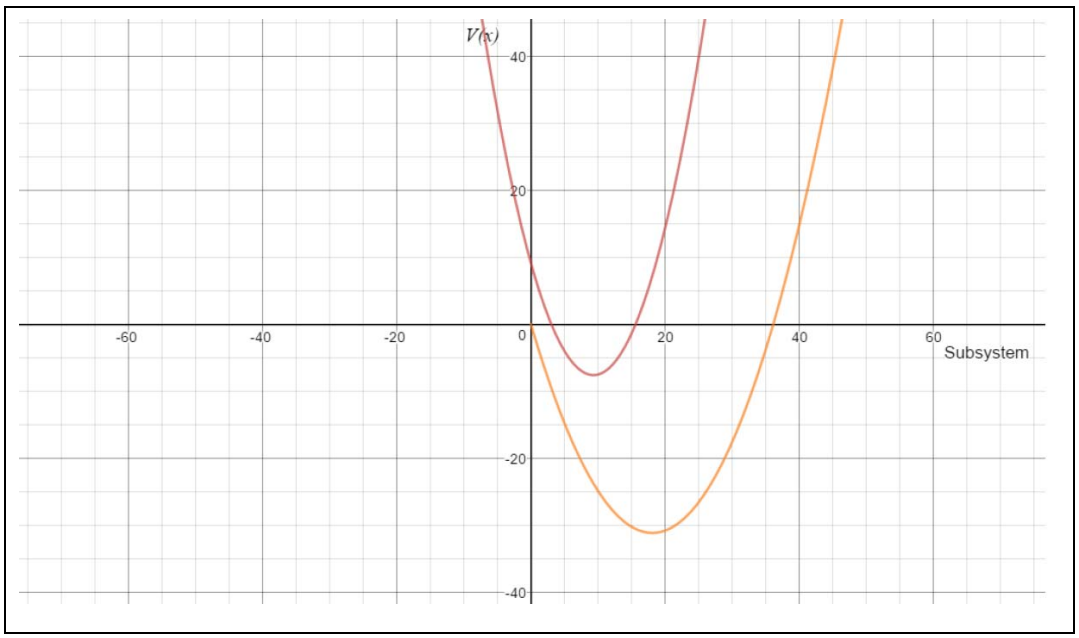
By analyzing the potential function with the simulation model, we have observed that it has a quartic shape similar to that of the cusp catastrophe model [22].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

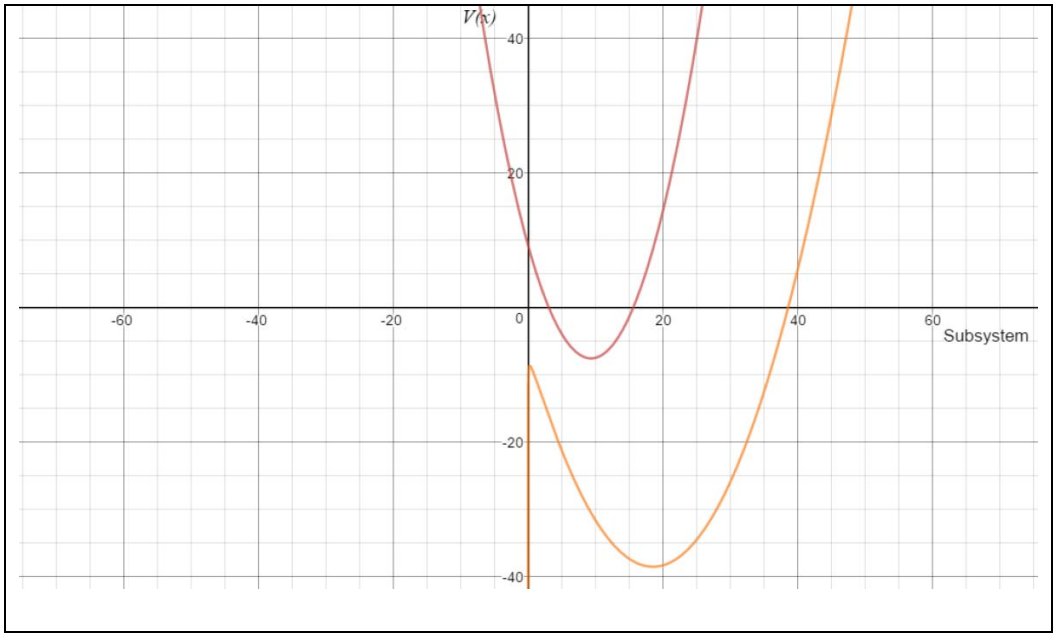
3.3.3 Study Of Emergence Of Catastrophe Indicators

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 [illegible] and the arrival rate with the behavior variable x as a number of working subsystems.

1. It has one minimum with one state.

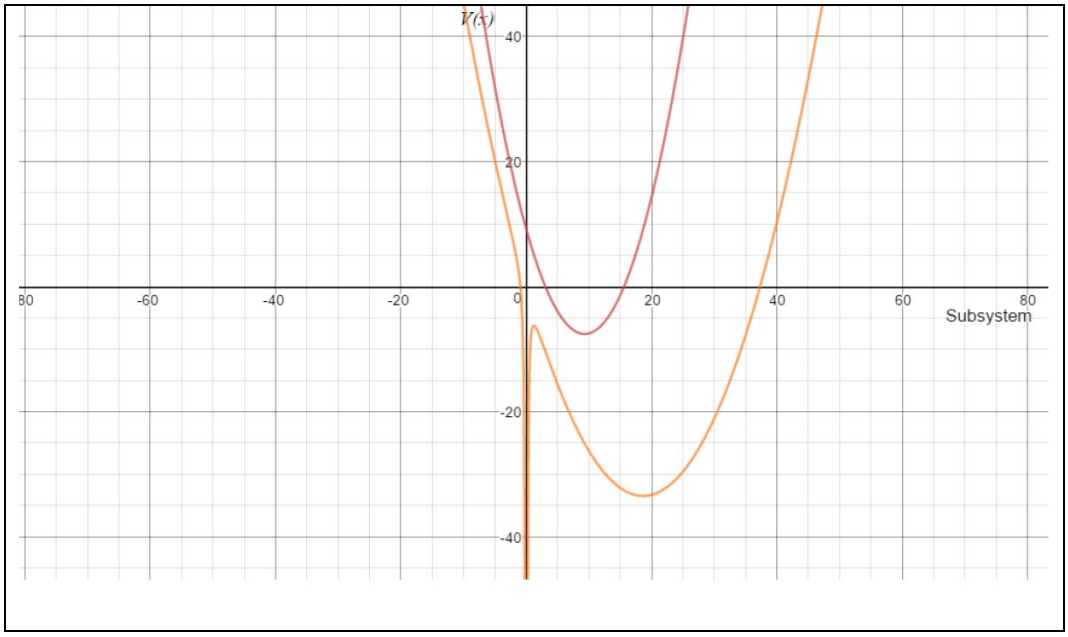


2. It has one minimum with a point of inflexion.



https://sites.google.com/site/ijcsis/
ISSN 1947-5500

3. It has two minimum and one maximum.



4. The new maximum disappears and leads to a new state.

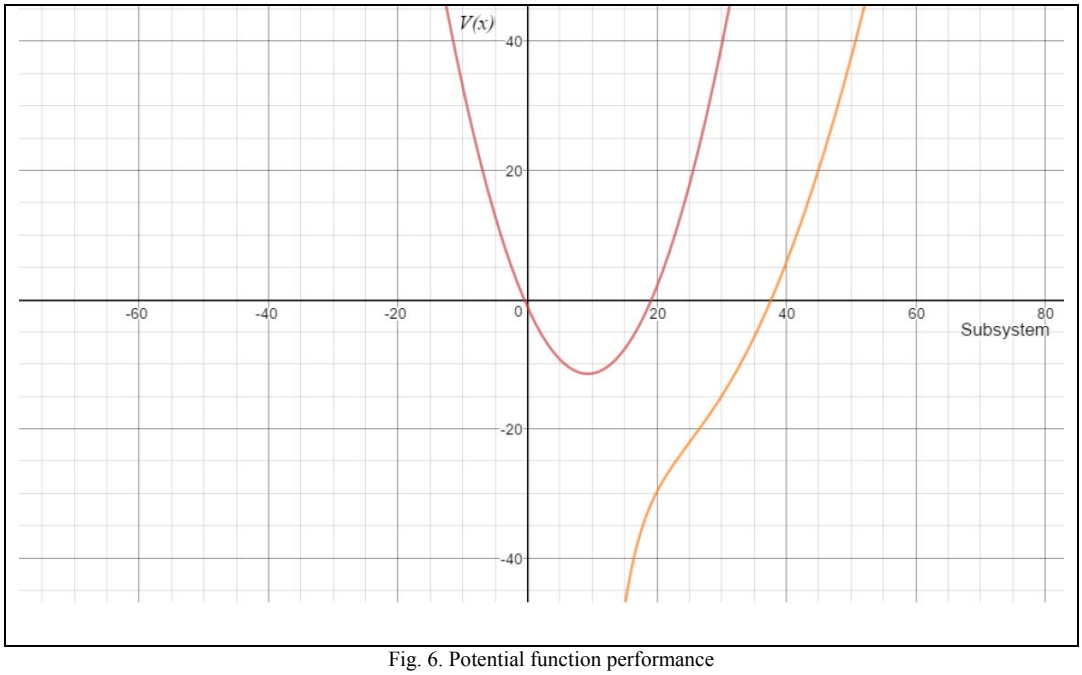


Fig. 6. Potential function performance

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 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.

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)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 works 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).

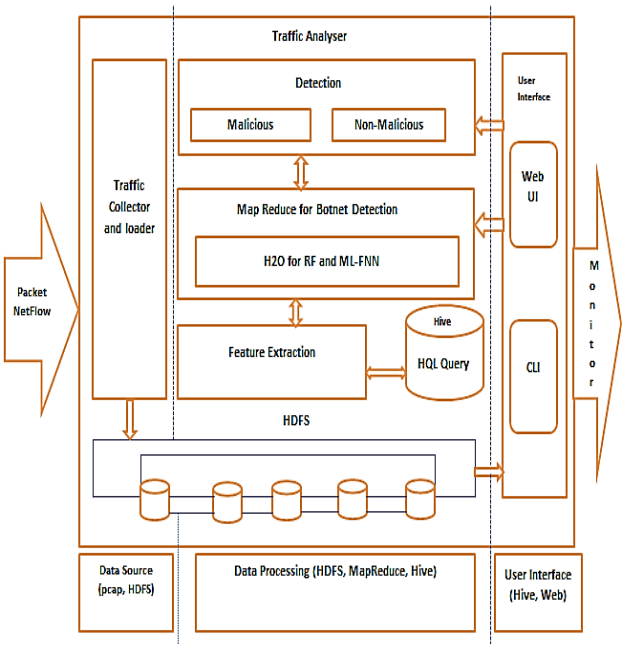


Figure.1 Scalable Framework for Botnet Detection

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:

a. Traffic Sniffer Module for packet pre-processing packets.

b. Feature Extraction Module for generating feature set.

c. 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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 exits 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)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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_bhlen	Total bytes used for headers in backward direction
Total_fhlen	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 <key, value> pairs where the input to framework is a set of <key, value> pairs and the output produced by the job is also another set of <key, value> pair.

(Input) <k1, v1> →map → <k2, v2> →combine →<k2, v2> →reduce→ <k3, v3> (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_bhlen
0.823	Std_bpktl
0.803	fpsh_cnt
0.7956	total_fhlen
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 connectinging 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 {Wi}1: N−1, where Wi represents the weight matrix of connecting layers i and i + 1 for a network of N layers.
- B is the collection {bi} 1: N−1, where bi denotes the column vector of biases for layer i + 1.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

H2O framework for multi-layer neural networks architectures are models of hierarchical feature extraction, it involves multiple levels of nonlinearity.

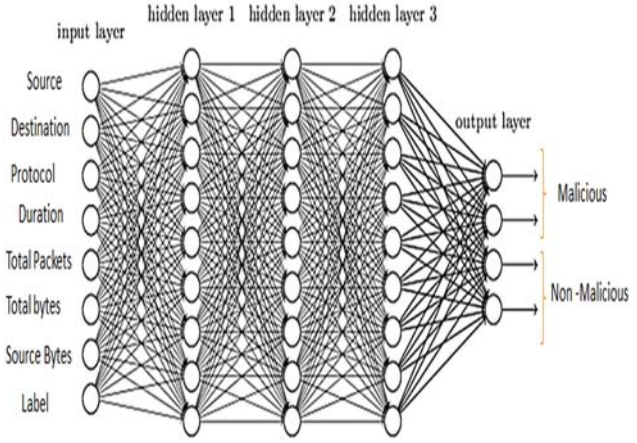


Figure.2 Multi-Layer Feed Forward Neural Network

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 ∇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 Tn, do in parallel:
   a. Get replica of the global model parameters Wn, Bn
   b. Select active subset Tna ⊂Tn (user-given number of samples per iteration)
   c. Separate Tna into Tnac by cores nc
   d. For cores nc on node n, do in parallel:
      i. Obtain training example i ∈Tnac
      ii. Update all weights wjk ∈ Wn, biases bjk ∈ Bn

   wjk:= wjk −α∂L (W, B) ∂wjk
   bjk:= bjk −α∂L (W, B) ∂bjk

   3.2 Assign W, B := Avgn Wn, Avgn Bn
   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 Wn, Bn. The Avgn 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, l1and 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.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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	Random Forest			Multi-Layer Feed Forward Neural Network		
	Precision	Recall	F-Score	Precision	Recall	F-Score
20k	0.9994	0.8410	0.9134	0.9873	0.9751	0.9874
40k	0.9996	0.8048	0.8917	0.9440	1.0	0.9712
60k	0.9996	0.7991	0.8882	1.0	0.9192	0.9579
80k	0.9996	0.7828	0.8780	1.0	0.9008	0.9478
1lk	0.9996	0.7396	0.8501	1.0	0.8804	0.9364
1.5l	0.9995	0.6894	0.8160	1.0	0.8589	0.9241
2l	0.9994	0.5809	0.7347	1.0	0.8389	0.9121

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.

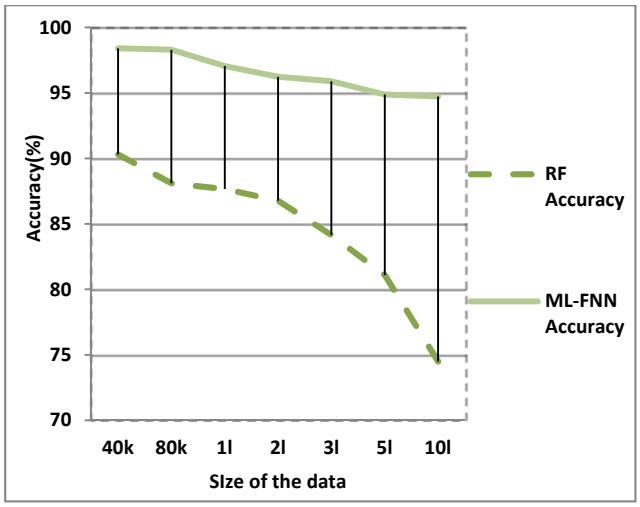


Figure.3 Comparison of accuracy

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

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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.
[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.
[9] F. Salcedo-Campos, J. Díaz-Verdejo, P. García-Teodoro, Segmental parameterisation and statistical modelling of e-mail headers for spam detection, Inf
[10] Reyhaneh Karimazad and Ahmad Faraahi. An Anomaly-Based Method for DDoS Attacks Detection using RBF Neural Networks. *IPCSIT* vol.ll (20ll) © (20ll) IACSIT Press, Singapore.
[11] G. Beydoun, G. Low, Generic modelling of security awareness in agent based systems, Inf. Sci. 239 (2013) 62–71.
[12] Y. Lee, Y. Lee, Toward scalable internet traffic measurement and analysis with hadoop, ACM SIGCOMM Comput. Commun. Rev. 43 (1) (2012) 5–13.
[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.
[17] W. Tarng, 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).
[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.
[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.com>.
[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.:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

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, CSVTU 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

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, Embeded 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 .

ISSN 1947 5500
http://sites.google.com/site/ijcsis/